Theoretical and Computational Strategies for the Study of Protein Folding Mechanisms

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Enrico Guarnera

aus
Italien

Promotionskomitee

Prof. Dr. Amedeo Caflisch (Vorsitz)
Prof. Dr. Ben Schuler

Zürich 2008

Die vorliegende Arbeit wurde von der Mathematisch-naturwissenschaftlichen Fakultät der Universität Zürich im Herbstsemester 2007 als Dissertation angenommen.

Promotionskomitee:
Prof. Dr. Amedeo Caflisch
Prof. Dr. Ben Schuler

Enrico Guarnera

Department of Biochemistry
University of Zürich
Winterthurerstr. 190
CH-8057 Zürich
Switzerland
guarnera@bioc.uzh.ch

Ordnung ist heutzutage meistens dort wo nichts ist.
Es ist eine Mangelerscheinung.
Bertold Brecht

Ogni limite ha una pazienza.
Totò

Alla mia famiglia, a Svava e... alla bella Vita.

Summary

This thesis focus on the development of computational tools to investigate the complex dynamics of protein folding mechanisms by means of molecular dynamics (MD) simulations. The study of the thermodynamic and kinetic properties of polypeptides requires the definition of appropriate probes. Defining a probe for a complex system often corresponds to provide a coarse-grained description of the ensemble of configurational microstates. That is especially the case when one has to extract relevant information from gigabytes of trajectories. In contrast to simple reactions of small and mainly rigid molecules the description of proteins is complicated because no natural order parameter exists. Here we investigate how the estimation of thermodynamic parameters depend on the choice of the coarse-grained or mesoscopic states. Particular attention is devoted to mesoscopic descriptions based on the "digitalization" of the configurational space using the secondary structure or the main chain torsional angles of protein microstates. The statistical mechanics of digital mesostates can be easily linked with information theory so that one can also quantify the usefulness of a description by evaluating its informational content. To study the folding kinetics a general framework is proposed based on Markov processes. The strategy is to use a minimal Markovian master equation to describe the dominant folding routes and the overal folding kinetics. The problem of the Markovianity of the MD trajectories is addressed and partially solved by an opportune redefinition of the mesostates that utilizes their causal connectivity. A simple algorithm is proposed to this purpose. The master equation technic is applied to equilibrium trajectories of the GSGS peptide, the triple stranded β-sheet peptide widely studied in our research group. This 20-residue peptide behaves as a two state folder although its free energy landscape presents many enthalpic and entropic basins different from the folded one. Among the results generated with the master equation analysis, it is clear that the basins of the unfolded states do not inter-convert on a time scale smaller than the folding time. Therefore, the folded basin plays the role of a network hub in the connectivity of the configurational space. The unfolded state seems not to be a region where the polypeptide is free to diffuse, rather it is pre-organized in basins from which independent and parallel folding routes depart.

Another topic investigated in this thesis was the study of simplified proteins by means of MD simulations. Simplified protein sequences were constructed by using an amino acid alphabet of solely three letters. We considered five proteins of two structural types: four full β-sheet proteins of respectively 20, 28, 36, 44 residues and an α/β protein of 56 residues. The α/β protein sequence is a simplified version of the B1 domain of protein G. It consists of alanines, threonines, and glycines at segments that in the folded structure are alpha-helical, beta-strand, and loop, respectively. The starting hypothesis is that for simple protein topologies (full β and α/β) low complexity amino acid alphabets are able to encode, although in diffuse manner, the overall structural properties of the folded states. This implies that the evolutionary patterns that generated the protein sequences were driven towards the specialization of protein functions rather than protein structures. Secondly, if the first hypothesis is true, can we simplify

protein sequences so that their folding mechanisms can be observed in a MD simulation? In equilibrium simulations we observed reversible folding for all the proteins studied. Most importantly the observed folded states corresponds to those the sequence design was aimed for. Notably the folded state of the simplified sequence for protein G, yielded the very same α/β structure of protein G. The folded states of each simplified sequence is marginally stable (about 1 kcal/mol) and highly accessible from the kinetic point of view. To date MD simulations of reversible folding have been reported only for structured peptides with less than about 20 residues, like alpha-helices, beta-hairpins and three-stranded beta-sheets, but not for globular proteins with a well-defined hydrophobic core. Here, reversible folding to the native structure of protein G is observed in more than 10-μs implicit solvent MD simulations at 330 K using a simplified sequence based on a three-residue alphabet. These results have twofold relevance: i) secondary structure propensities alone are able to encode the folded conformation; ii) the 20 different types of side chains have been selected by natural evolution for optimizing protein function (and preventing pathological aggregation) but not to increase folding speed and/or stability.

Zusammenfassung

Diese Dissertation konzentriert sich auf die Entwicklung von rechnergestützten Methoden, um die komplexen Mechanismen der Proteinfaltung mittels Molekulardynamik (MD)-Simulationen zu untersuchen. Die Untersuchung von thermodynamischen und kinetischen Eigenschaften von Polypeptiden erfordert die Definition passender Observablen. Eine Observable für ein komplexes System zu definieren bedeutet oft eine "coarse-grained" Beschreibung des Ensembles von Konfigurationsmikrozuständen zu erstellen. Dies wird besonders wichtig, wenn man relevante Informationen aus Trajektorien von mehreren Gigabytes extrahieren muss. Im Gegensatz zu simplen Reaktionen von kleinen, meist rigiden Molekülen, ist die Beschreibung von Proteinen komplizierter, weil kein natürlicher Ordnungsparameter existiert. In dieser Arbeit untersuchen wir, wie die Abschätzung von thermodynamischen Parametern abhängt von der Wahl der Systembeschreibung oder der Definition der mesoskopischen Zustände. Besondere Beachtung wurde der mesoskopischen Beschreibungen gewidmet, die auf der Digitalisierung des Konfigurationsraums mittels Sekundärstruktur oder der wichtigsten Torsionswinkel der Protein-Mikrozustände basiert. Die statischische Beschreibung der diskreten Mesozustände kann einfach mit der Informationstheorie verbunden werden. Damit kann die Beschreibung quantifiziert werden, indem man deren informationstheoretischen Inhalt bestimmt. Um die Kinetik der Faltung zu untersuchen, wird ein allgemeines Modell basierend auf Markov-Prozessen vorgestellt. Die Grundidee besteht darin, eine Markov-Mastergleichung zu benutzen, um die wichtigen Faltungswege und die globalen Faltungskinetiken zu beschreiben. Das Problem der Markov-Eigenschaft der MD Trajektorien wird behandelt und eine Teillösung in Form eine Neudefinition der Mesozustände, die deren kausale Konnektivität benutzt , präsentiert. Dazu wird ein einfacher Algorithmus vorgestellt. Die Methode basierend auf Mastergleichungen wird auf Gleichgewichtstrajektorien des GSGS-Peptids angewandt, ein kleines Modell-β-Faltblatt-Peptid, welches in unserer Gruppe untersucht wird. Dieses Peptid bestehend aus 20 Aminosäuren verhält sich gemäss des Zweizustandsmodels, obwohl die Energielandschaft viele enthalpische und entropische Becken aufweist, die sich vom Becken des gefalteten Zustands unterscheiden. Ein Resultat aus der Analyse mittels Mastergleichungen besteht darin, dass die ungefalteten Zustände in unterschiedlichen Becken nicht während einer Zeit kleiner als der Faltungszeit ineinander übergehen. Daher nimmt das Becken der gefalteten Zustände die Rolle einer Drehscheibe für die Verbindungen zwischen unterschiedlicher Regionen des Konfigurationsraums ein. Nach unseren Resultaten erscheint der ungefaltete Zustand nicht als Region, in der das Polypeptid frei diffundieren kann, sondern die ungefalteten Zustände sind strukturiert in Becken aus denen unabhängige, parallele Wege zum gefalteten Protein ausgehen.

Eine weitere Thematik dieser Dissertation ist die Untersuchung von vereinfachten Proteinen mittels MD Simulationen. Vereinfachte Proteinsequenzen wurden konstruiert anhand eines Aminosäure-Alphabets mit nur drei Buchstaben. Wir betrachten fünf Proteine mit zwei strukturellen Motiven: vier β-Faltblatt-Proteine mit jeweils 20, 28, 36 und 44 Aminosäuren und ein α/β Protein mit 56 Aminosäuren.

Die α/β Proteinsequenz ist eine vereinfachte Version der B1 Domäne von Protein G. Sie besteht aus Alaninen, Threoninen und Glycinen in Abschnitten, die in der Wildtyp-Struktur jeweils α-Helix, β-Faltblatt and einen Loop bilden. Die Ausgangshypothese ist, dass für gewisse Proteintopologien ein Aminosäure-Alphabet von geringer Komplexität in der Lage ist, die globalen strukturellen Eigenschaften des gefalteten Zustands approximativ zu kodieren. Das bedeutet, dass die Evolution Proteinsequenzen in Richtung der Spezialisierung von Funktion anstelle von Struktur vorangetrieben hat. Ausserdem unter Annahme der ersten Hypothese stellt sich die Frage, wie können wir Proteinsequenzen vereinfachen, so dass ihr Faltungsmechanismus in MD Simulationen beobachtet werden kann? In Gleichgewichts-Simulationen wurden Faltungs-/Entfaltungsereignisse für alle untersuchten Proteine beobachtet. Vor allem entsprechen die beobachteten gefalteten Zustände denen, die der angestrebten, strukturellen Vorgabe entsprechen. Besonders der gefaltete Zustand der vereinfachten Sequenz für Protein G ergab nahezu dieselbe α/β Struktur von Protein G. Die gefalteten Zustände jeder vereinfachten Sequenz sind nur gering stabil (etwa 1 kcal/mol) und sehr zugänglich in kinetischer Hinsicht. Zur Zeit sind nur MD Simulationen mit reversibler Faltung von strukturierten Peptiden mit weniger als 20 Aminosäuren bekannt, wie beispielsweise α-Helices, β-Haarnadeln und 3-strängige β-Faltblätter, nicht aber für globuläre Proteine mit stark ausgeformtem hydrophoben Kern. Hier wurde die reversible Faltung zur native Struktur von Protein G beobachtet in MD Simulationen mit mehr als 10 μs Länge mit impliziter Behandlung des Lösungsmittels bei 330 K , wobei die Sequenz auf ein drei-Aminosäure-Alphabet vereinfacht wurde. Diese Ergebnisse haben zweifache Bedeutung: i) Eine Auswahl entsprechend der Sekundärstruktur-Tendenzen der Aminosäuren genügt um die gefaltete Konformation zu reproduzieren; ii) die 20 verschiedenen Seitenketten wurden selektioniert durch natürliche Evolution für Proteinfunktion (und Prävention von pathologischer Aggregation), aber nicht in Hinblick auf Faltungsgeschwindigkeit und/oder Proteinstabilität.

Contents

Summary II

Zusammenfassung III

Contents V

1 Introduction 1

1.1 Where is the problem in protein folding? . . . 1
1.2 What models for folding? . . . 4
1.3 Reaction coordinates and the importance of the description in complex systems . . . 7
1.4 Simplified protein sequences as a strategy to study folding . . . 11
1.5 The structure of this thesis . . . 14

2 Protein folding mechanisms from MD simulations 17

2.1 Introduction: reducing the complexity . . . 17
2.2 Methods based on structural similarity . . . 18
2.3 Symbolization of the conformation space . . . 20
2.3.1 Strings of native contacts . . . 20
2.3.2 Strings of secondary structure . . . 21
2.3.3 Strings of rotational states . . . 21
2.4 Thermodynamics of the coarse graining . . . 21
2.4.1 Entropy, information and order parameters . . . 22
2.4.2 Disorder and complexity of the ensemble of mesostates . . . 33
2.4.3 Non-convergence of the mesostates . . . 35
2.4.4 Rank ordered distributions and density of mesostates . . . 36
2.4.5 Structural interpretation of the mesostates . . . 41
2.5 The organization of an ensemble of strings . . . 43
2.5.1 Entropy and disorder of strings . . . 43
2.5.2 String hierarchy and configurational hierarchies . . . 47
2.5.3 Convergence of the entropy . . . 49
2.6 Folding kinetics in the space of mesostates . . . 51
2.6.1 First passage times . . . 51
2.6.2 Mean first passage times . . . 54
2.6.3 Folding kinetics hierarchy . . . 58
2.6.4 Local rates and the Zwanzig model . . . 59

2.7 Causal mesostates and conformational master equation . . . 62
2.7.1 Markov approximation . . . 63
2.7.2 Estimating a stochastic matrix . . . 64
2.7.3 Causal grouping of the mesostates . . . 67
2.7.4 A Markov chain on the causal grouped mesostates . . . 70
2.7.5 Network representation of the transition matrix . . . 73
2.7.6 MFPTs from a Markov chain . . . 77
2.8 Conclusions . . . 80

3 Simulations studies on simplified protein sequences 83
3.1 Introduction . . . 83
3.2 Simplifying strategies . . . 85
3.3 Methods . . . 88
3.3.1 Sequences . . . 88
3.3.2 Molecular dynamics simulations . . . 89
3.3.3 Description of the configurational space . . . 90
3.4 Results . . . 90
3.4.1 Statistical thermodynamics of the configurational spaces . . . 90
3.4.2 Organization of the configurational space of "primitive" proteins . . . 100
3.4.3 Folding kinetics: first passage time analysis . . . 104
3.4.4 Folding kinetics: pathways hierarchy . . . 107
3.4.5 Folding kinetics: Markovian dynamics and causal grouping . . . 109
3.5 Conclusions . . . 115

4 How does a simplified-sequence protein fold?
(Submitted manuscript) **117**

5 Estimation of protein folding probability from equilibrium simulations
(*Journal of Chemical Physics* (2005) **122**, 184901) **141**

6 Pathways and intermediates of amyloid fibril formation
(*Journal of Molecular Biology* (2007), **379**, 917-924) **147**

7 Conclusions and outlook 163

Bibliography 168

List of figures 192

List of tables 194

Acknowledgments 195

1 Introduction

1.1 Where is the problem in protein folding?

The studies on the renaturation of the RNase in solution carried out by Anfinsen during the 50s which led him to the nobel price in 1973 [Anfinsen, 1973], can be seen as the historical foundation of protein folding as an autonomous branch of the molecular biology. Since then the knowledge on the protein folding problem is amazingly increased in both the experimental and the theoretical fields. Knowledge on the biochemistry of proteins has dramatically expanded too in the last 30 years. Today we know certainly better than yesterday what is the role played by proteins in the living matter. As biological macromolecules the primary purpose of proteins is to perform functions within the cell [Alberts et al., 1998]. Functions performed are essentially of three types: enzymatic, cell signaling or signal transduction and structural. The enzymatic activity is the best known role played by proteins which are, such as enzymes, in charge for the catalysis of chemical reactions. Enzymes carry out most of the chemical reactions involved in the metabolism. Several enzymes can work together in a specific order, creating *metabolic networks*. In cell signaling some kind of proteins are responsible of the transmission of signals from a cell where they were synthesized to other cells of the organisms. Others are membrane proteins that act as receptors whose main function is to bind a signaling molecule and induce a biochemical response in the cell, or the ligand transport proteins, such as hemoglobin for instance, which bind particularly small biomolecules and transport them to other locations in the organism. Structural proteins provide rigidity to the fluid-like biological matter: examples are fibrous proteins, collagen and motor proteins. To use a computer science language, as DNA represents the program in which it is encoded all the design of an organism, proteins act as the "executors" of the DNA prescriptions. Proteins are then the actors which make the "dirty job" of making life functioning, or to put it in a philosophical manner, they accomplish to what in Aristotelianism is called teleologic activity, namely the purpose of realizing what is planned in the DNA. This metaphor is the basis of the so called *central dogma of the molecular biology* which states that the information flow is transmitted in cascade from DNA to proteins and never goes back [Crick, 1970]. In all cases the relationships between protein shapes and functions are intimately close and in many of the cases proteins fold into a specific three dimensional structure to accomplish their function. That is what is commonly acknowledged as the protein structure/function paradigm, which reads that in order to be functional a protein has to assume an ordered three dimensional structure under physiological conditions. Initiated and supported by the pioneer works of Mirsky and Pauling [Mirsky and Pauling, 1936], and Wu's [Wu, 1931], this hypothesis led to an extraordinary production of protein structures using both X-ray and NMR experimental technics [Berman et al., 2003]. Anfinsen was the first calling "thermodynamic hypothesis" the structure/function paradigm, namely the conjecture according to which the native state of a protein corresponds to a global minima of the

free energy [Anfinsen, 1973]. Logic consequence of this hypothesis is that all the structural information about the native state of the protein must be completely determined by its sequence of amino acids, which is the final result of the biological evolution. Thus it is clear that the protein folding problem according to Anfinsen's view is nothing else than this: given the sequence of a protein to determine its unique three dimensional functional structure. If a unique structure is required for a protein in order to be functional then, since a polypeptide can in principle assume an astronomical number of different configurations, an unavoidable "search problem" is also implied. In 1968, in his famous one page article, Levinthal already guessed the deepness of the search problem by proposing the paradox that has his name [Levinthal, 1968]. Stated simply the paradox reads as follow: how can an unfolded polypeptide chain, that is free to sample the vastness of configurational space, find the native conformation in a biological time after a shift to physiological conditions? Since there are mainly two elementary degrees of freedom in the peptide unit, the torsional angles ϕ and ψ whose allowed values depends on the stereo-chemistry of the amino acids [Ramachandran et al., 1963], assuming four possible configurations per peptide unit (a pair of ϕ and ψ angles), then for a 100 residues polypeptide there are $4^{100} \sim 10^{60}$ possible configurations. With a picosecond time scale for bond rotations the time needed to sample all of them would be of the order of 10^{40} years which is enormously greater than the about 10^9 years for the estimated age of the universe. Thus quoting Levinthal, "if the final folded state turned out to be the one of lowest configurational energy, it would be a consequence of biological evolution and not of physical chemistry", so that the paradox is not a paradox at all and folding must be a consequence of a direct, biased and perhaps "intelligent" search. The question is how that is possible. Anfinsen's argument according to which native proteins are in their global free energy minima, constitutes a qualitative answer to the Levinthal quest. At the time of Anfinsen's Nobel prize the denatured state of proteins was thought to be essentially ruled by the random coil statistics, namely without any structural content and following the Flory hypothesis. The isolated-pair Flory hypothesis states that under denaturing conditions the cross correlation between the chain units is negligible, i.e. each (ϕ, ψ) pair is sterically independent, which is almost equivalent to say that the configurational ensemble in the denatured state is structureless [Flory, 1969, Flory, 1974]. If we consider a polypeptide chain of n residue in its denatured state and assume it satisfies the random coil statistics, then if λ rotational states per (ϕ, ψ) torsional pair are allowed, the total main chain configurational entropy can be estimated as

$$S_{\rm denat} \sim k_B \ln \lambda^n = n\, k_B T \ln \lambda \tag{1.1}$$

which means that under the random coil hypothesis any configuration is equally accessible. As soon as refolding conditions are restored the protein spontaneously picks a specific configuration, the native one, out of all the possible ones, which means a dramatic entropic collapse. If the native state is configurationally well defined the configurational entropy is $S_{\rm nat} \sim 0$ so that the protein entropy loss due to refolding is simply $\Delta S_{\rm fold} = S_{\rm nat} - S_{\rm denat} \sim -n\, k_B \ln \lambda$ which has a negative sign. Such an entropic crisis must be then sustained by an enthalpy gain in favor of the native state in such a way it can be in a global free energy minima, namely

$$\Delta G_{\rm fold} = \Delta E_{\rm fold} - T\Delta S_{\rm fold} < 0 \tag{1.2}$$

which means $\Delta E_{\rm fold} < T\Delta S_{\rm fold}$ with $\Delta E_{\rm fold}$ the enthalpy gain for the protein under refolding conditions. If we take $\lambda = 3$, three rotational states per chain unit, and consider a 100 residue chain at a

physiological temperature of 300 K, the enthalpy gain for folding should be $\Delta E_{\rm fold} \lesssim -82$ kcal/mol which is a fairly enough large value considering that most of globular proteins are only marginally stable, regardless of size and activity with $\Delta G_{\rm fold}$ in the range of -2/-10 kcal/mol [Privalov and Khechinashvili, 1974, Makhatadze and Privalov, 1995, Taverna and Goldstein, 2002, Baldwin, 2007]. Thus most of the enthalpy gain ought to be viewed as a compensatory effect of the entropy loss that alone drives the folding process. This enthalpy driven folding picture has led the scene in the development of the theories of protein folding. An example of that is the development of native centric models such as the Go models [Go, 1983]. In these models the native structure is considered as the target state of a dynamical process in which only the native interactions are favorable and drive folding until, by definition, the folded configuration turns out to lie in a minima of the potential energy (and thus enthalpy). Despite the great evolution of these kind of models, also in the direction of a more realistic treatment of the non-native interactions [Karanicolas and Brooks, 2003], their limits are nevertheless evident [Cavalli et al., 2005] since they essentially play the role of descriptive, rather than predictive models.

In the compute of the enthalpy gain due to folding one has to take into account four main forces [Dill, 1990a]: the van der Waals attractions which depends on the stereo-chemistry of the amino acids and arise from interactions among fixed or induced dipoles; the hydrogen bond which occurs when an hydrogen atom is shared between two electronegative atoms with a strength in the range 2-10 kcal/mol; the electrostatic forces which are responsible for the long range interactions; the hydrophobic effect which gives account of the aversion for water of the non-polar amino acids. The contributions to the free energy of folding due to the solvent are generally divided in two parts, the polar and non-polar, the former affecting the hydrophilic residues and the latter affecting those hydrophobic. The non-polar contribution at physiological temperature is essentially entropy driven. The transfer of non-polar species in water at 300 K is not opposed by the enthalpy and conversely is favored by entropy because the waters prefer to hydrogen bonding to other waters instead to make hydrogen bonds with the non-polar species [Dill, 1990a, Dill, 1990b]. The phenomena essentially corresponds to of phase separation between solute and solvent which allows the protein to have access to a greater number of configurations, that is properly known to be an entropy driven phase transition [Frenkel, 1999].

Determining how a given polypeptidic sequence folds to a stable protein with a well defined structure is one of the great challenges of nowadays theoretical biophysics [Onuchic et al., 1997, Pande et al., 2000, Dinner et al., 2000, Thirumalai et al., 2002, Fersht and Daggett, 2002]. Computer simulations are a powerful technique to investigate the structure and the dynamics of proteins at atomic resolution. In principle, computer simulations could determine the complete free energy surface of proteins and thus the native, folded conformation in correspondence of the global minimum of the free energy as well as barriers height and rates. In practice, however, equilibrium properties of even small-size proteins with 50 or more amino acids cannot be determined in this way as this would in general require the simulation of long trajectories in the μs to ms range (see e.g. the Protein Folding Database http://pfd.med.monash.edu.au) that cannot be calculated with current computers. While currently available sequence-based force-fields might not be able to properly fold a protein to its native structure using molecular dynamics (MD) simulations, the past few years computationally efficient all-atom implicit solvent models have made it possible to describe the reversible folding of some peptides and miniproteins [Schaefer et al., 1998, Gnanakaran et al., 2003, Caflisch and Paci, 2004].

1.2 What models for folding?

What kind of mechanisms can explain the great entropy loss of the denatured state upon folding? Before to briefly explore the theoretical suggestions that in the last decades have been proposed, three well established facts needs to be reminded. The first is the discovery of the "molten globular state" which plays the role of a major kinetic intermediate near the start of the folding pathway [Kuwajima, 1989, Ptitsyn et al., 1990, Ptitsyn, 1992]. Discovered around the 80s in a protein environment under mild denaturing conditions (low pH), it is a thermodynamic phase structurally characterized by a native-like secondary structure associated to a fluctuating backbone [Vassilenko and Uversky, 2002], a lack of specific tertiary contacts (liquid like) with typically the presence of a loosely packed hydrophobic core [Ptitsyn and Uversky, 1994]. Experimental data clearly indicates that in the folding reaction the transitions form denatured to molten globule and from molten globule to native resemble first order phase transitions [Koshiba et al., 2001], that are characterized by a very fast time scale in the former (about 10^{-3} s) and a slow time scale in the latter (the folding time) [Arai and Kuwajima, 1996]. Although not completely unanimous (see for example [Creighton, 1997]), the molten globular phase of proteins is widely accepted to play a key role in protein folding.

The second important fact is the reassessment of the denatured/unfolded state in light of the inconsistency of the random coil hypothesis. In the last years it has been rooted the idea that the deep comprehension of protein folding pass through a complete understanding of the denatured state [van Gunsteren et al., 2001]. For an increasing number of proteins it has been experimentally shown that the denatured/unfolded state is structured: remarkable examples are the works of Shortle who showed that, even under strong denaturing conditions, such as 6M GuHC1 and 9M urea, a residual structure may survive that, in some cases, correlates to a native-like signal [Dill and Shortle, 1991, Shortle and Ackerman, 2001]. Notably under milder or physiological conditions, the denatured states of most proteins appear to be highly compact with a high secondary structure content and moreover, as mutational analysis has suggested, it plays a key role in protein stability [Shortle, 1996]. Yet, studies on the fast kinetics formation of loop/turn in unfolded polypeptides [Fierz et al., 2007] and the residual dipolar couplings (RDCs) observed by NMR spectroscopy on unfolded model peptides, showed a strong orientational preference of the amino acids [Dames et al., 2006] and suggest a pre-organized view of the denatured/unfolded as outlined by Rose and collaborators [Rose et al., 2006]. A pre-organized unfolded state is largely dominated by the local interactions between adjacent residues whose steric and hydrophobic interactions drive the local configurational preferences. Thus, such a residual structure in the denatured state consequently implies the questioning of the Flory's assumptions on the isolated pair hypothesis [Pappu et al., 2000].

The last fact we would like to recall here is the increasing evidence of the role of the intrinsically disordered or natively unfolded proteins. Natively unfolded or intrinsically unstructured proteins constitute a unique group of the protein universe. In particular their evolutionary persistence is a strong indication of their relevant biological role. These proteins show a low level of ordered secondary structure and no tightly packed core. They are very flexible, but can assume relatively rigid configuration under binding with ligands. In comparison with the globular proteins, natively unfolded proteins are very sensitive to the changes in environment [Uversky, 2002]. About the 10 % of proteins are predicted to be fully disordered, while at least the 40 % of eukaryotic proteins have at least one long (>50 residues) disor-

dered region [Tompa, 2002]. It has been suggested that the intrinsically unstructured proteins might be the rule instead the exception on how the protein universe is organized. In particular this suggestion has lead to the proposal that proteins might generally exists in a "trinity" of states: the ordered state, the molten globule, and the disordered, all of the three having a functional role depending on the environment [Dunker et al., 2001]. Thus, according to these new view the classical sequence/structure paradigm for proteins would hold only for a tight subclass of the protein universe. Natively unfolded proteins configure themselves as an advanced stage of the evolution of proteins for their multiple ability to perform functions in metabolic networks [Gavin et al., 2002].

In light of the three established facts recalled above, an ideal theoretical framework of folding, or more generally protein behavior, should be a minimal compromise able to include all the diverse configurational features of proteins. In the last decades the debate between experimentalists and theoreticians has focused on globular proteins, in particular whether they reach their global energy minimum in a pathway-independent manner under thermodynamic control (the thermodynamic hypothesis), or whether they follow a specific pathway to a possibly local minimum under kinetic control. Experiments have clearly suggested that some small monodomain proteins obey the thermodynamic hypothesis [Kim and Baldwin, 1990, Dill, 1990a]. However, there are many examples of proteins whose functional native state is metastable while more stable inactive conformations are avoided: instances are the plasminogen activator inhibitor (PAI-1) [Berkenpas et al., 1995] as well as other members of the serpin family [Cabrita and Bottomley, 2004]. Another example of process under kinetic control is the protein misfolding, that in many diseases such as Alzheimer's, Creutzfeldt-Jakob's, bovine spongiform encephalopathy, and Parkinson's, is held to be the cause of ordered aggregation. Misfolded states, such as kinetic traps, intermediates or competitive states of lower energy than the native one, are able to escape to the control mechanisms within the cell leading first to the malfunctioning of the cellular metabolism and then to the disease [Dobson, 1999, Dobson, 2003]. The models that have been suggested to explain the high speed of protein folding are essentially three, as depicted in figure 1.1. In the hierarchical model (figure 1.1 (A)) [Ptitsyn and Rashin, 1975, Kim and Baldwin, 1982, Kim and Baldwin, 1990, Baldwin and Rose, 1999b, Baldwin and Rose, 1999a, Rose et al., 2006, Ozkan et al., 2007], folding is initiated by the formation of the elements of secondary structure already in the unfolded state regardless of the tertiary contacts. Essentially the local propensities of adjacent amino acids drive the formation of backbone-backbone hydrogen bonds which are responsible for the construction of the secondary structure. This step of local arrangement can be very fast depending on the strength of the local preferences of the residues. The scenario suggested by Rose and coworkers of a pre-organized (or pre-sculped) unfolded state and the experimental results on the residual structure in the denatured state, seem to support such a view of the early stages of folding. An image of pre-organization can be a denatured state shaped like a swiss cheese, where the holes correspond to gateways for the folding channels which, connected like in a branched tree, drive the system to its folded state. Each hole in this picture corresponds to a local region of the polypeptidic chain from which folding initiates. Moreover, the Ising like model firstly proposed by Zwanzig at the beginning of the 90s, already showed that strong propensities of the amino acids to form ordered local structures can be an extraordinary bias to easily overcoming the Levithal's paradox and providing biological folding times [Zwanzig, 1995, Zwanzig et al., 1992]. Once the local elements of secondary structure are formed they can get assembled by means of a diffusion-collision mechanism [Karplus and Weaver, 1979, Karplus and Weaver, 1994]. The elements of secondary struc-

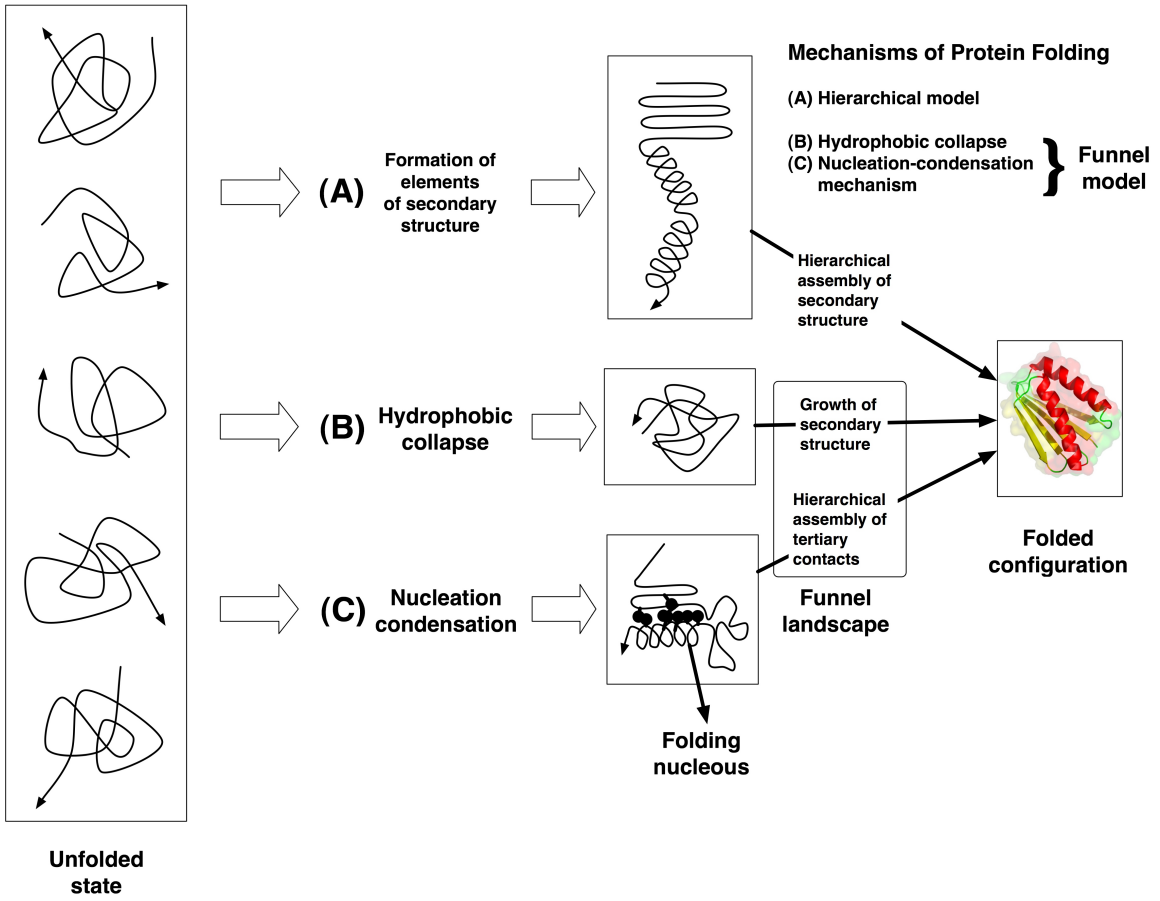

Figure 1.1: Possible mechanism for protein folding. (A) The hierarchical model [Ptitsyn and Rashin, 1975, Kim and Baldwin, 1982, Kim and Baldwin, 1990, Baldwin and Rose, 1999b, Baldwin and Rose, 1999a, Rose et al., 2006, Ozkan et al., 2007]. Protein folding is thought to start with the formation of elements of secondary structure independently of tertiary structure, or at least before tertiary structure is locked in place. These elements then assemble into the tightly packed native tertiary structure by means of a diffusion-collision (or framework) mechanism [Karplus and Weaver, 1979, Karplus and Weaver, 1994]. (B) Hydrophobic collapse model for folding [Rackovsky and Scheraga, 1977, Dill, 1985, Dill, 1990b]. The initial event of the reaction is thought to be a relatively uniform collapse of the protein molecule, mainly driven by a phase separation given by the hydrophobic effect. Stable secondary structure starts to grow only from the collapsed state. (C) Nucleation-condensation mechanism [Fersht, 1995, Itzhaki et al., 1995, Shakhnovich et al., 1996, Fersht, 1997, Fersht, 1999, Kiefhaber et al., 1997]. Early formation of a diffuse protein-folding nucleus catalyses further folding. The nucleus primarily consists of a few adjacent residues which have some correct secondary structure interactions, but it is stable only in the presence of further approximately correct tertiary structure interactions. Both mechanisms (B) and (C) at final stage of folding are compatible with the funnel model [Bryngelson et al., 1995, Dill and Chan, 1997]. Figure adapted from [Nolting and Andert, 2000].

ture (helices or β-hairpins) diffuse in the space so that favorable tertiary interactions can be found due to entropic effects as a consequence of the hydrophobic effect. It is clear that such a paradigm of folding implies parallel routes and a not necessarily a structurally defined transition state. According to the hydrophobic collapse model (figure 1.1 (B)) [Rackovsky and Scheraga, 1977, Dill, 1985, Dill, 1990b] the initial event of folding is an overall collapse of the chain so that secondary structure can grow within a compact configurational space. On the other hand in the nucleation-condensation mechanism [Fer-

sht, 1995, Itzhaki et al., 1995, Shakhnovich et al., 1996, Fersht, 1997, Fersht, 1999, Kiefhaber et al., 1997] (figure 1.1 (C)) a critical diffuse folding nucleus involving a low number of close residues in a critical network of tertiary contacts [Vendruscolo et al., 2001], catalyzes further folding and the growth of secondary structure under the pressure of the hydrophobic effect. In these two models the formation of a key nucleus implies the existence of a structurally well defined transition state and, consequently, less heterogeneous folding pathways. Moreover, these two models are both compatible with the framework of funneled energy landscapes [Bryngelson et al., 1995, Dill and Chan, 1997]: the critical nucleus is the rate limiting step of the folding reaction after which the growth of favorable native interactions, dramatically decreases the enthalpy of the protein compensating the entropy loss and eventually overtaking it. This view is also considered supported by the phi-value analysis which, as an experimental protein engineering method, has been widely used to study the structural features of the folding transition state in small protein domains that mainly fold in a two-state manner (all-or-none) [Fersht, 1999, Mayor et al., 2003].

1.3 Reaction coordinates and the importance of the description in complex systems

Proteins are perfect example of complex system inasmuch as they are characterized by many degrees of freedom which interact in a non trivial manner. To quote the words of the mathematician and Nobel price Herbert Simon in his "The architecture of complexity" [Simon, 1962]: "*Roughly, by a complex system it is meant one made up of a large number of parts that interact in a nonsimple way. In such systems the whole is more than the sum of the parts, not in an ultimate, metaphysical sense but in the important pragmatic sense that, given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole.*"

Closely related to the Simon's statement there is the problem of the *description* for complex systems and specifically for proteins. The investigation of the problem of the description of the configurational space of proteins and the development of methods on this subject is one of the main topics of this thesis work. Our tool of investigation is basically rooted in the use of MD simulations. The trajectories generated in simulations are considered here as the outcome of a special kind of experiment in which full access to the complex dynamics of the system is possible. In simulations based on an atomistic description of the protein molecules, the full access to all the degrees of freedom of the system is in principle feasible. However, given the multidimensional character of the configurational space of proteins, even if a foldable polypeptide clearly admits at least two macro-phases, a folded and an unfolded one, it is not trivial to define a quantity able to distinguish among all the configurational states. If we focus our attention on foldable proteins, folding reaction is a transition from disorder (or partial order) to full order. In terms of free energy the transition goes from a macro-phase (the unfolded one) in which the free energy is largely dominated by high entropy (large TS_{unfold}) to a macro-phase (the folded one) whose free energy is characterized by low enthalpy ($\Delta E_{\mathrm{fold}} \ll 0$). The description problem arises when the investigator looks for quantities able to distinguish these macro-phases on both thermodynamic and kinetic sides. These quantities when referred to thermodynamics are called "order parameters" or "reaction coordinates" when a kinetic context is needed. Ideally order parameters should be able to discern

among the thermally stable phases by approximating the free energy function of the system. On the other hand, ideal reaction coordinates should give full account of the free energy barriers that the system needs to cross to reach its equilibrium. The term macro-phases is appropriated inasmuch as a phase is the average over an ensemble of microstates characterizing the phase-space of the system. Microstates are states fully characterized by the molecule spatial degrees of freedom and momentums. Less general microstates are the configurational microstates which are characterized by only the spatial degrees of freedom only. The probability to have a certain microstate in phase space at a given fixed temperature is defined by the normalized Gibbs-Boltzmann probability density

$$\rho(\Gamma) = \frac{1}{Z} e^{-\beta E(\Gamma)} \tag{1.3}$$

with Γ the element of the phase space, Z the canonical partition function, $\beta = 1/k_B T$. The Gibbs-Boltzmann probability density characterizes the whole thermodynamics of the system, in particular its entropy[1]

$$S = k_B \int \rho(\Gamma) \ln \rho(\Gamma) d\Gamma \tag{1.4}$$

and free energy

$$G = \overline{E} - TS \tag{1.5}$$

with $\overline{E}$ the mean internal energy of the system or its enthalpy[2]. Let us first focusing on order parameters. Defining an order parameter for folding corresponds to define a probe. To use a language more familiar to experimentalists, a probe means to design a measuring device (CD spectra, fluorescence, NMR spectroscopy, light scattering, single molecule spectroscopy, etc.) that leads to a complete set of outcomes. If we imagine these outcomes discrete and finite in number, for instance $i = 1, \cdots, N$, so that a probability or an intensity P_i can be measured, then we would have the normalization $\sum_{i=1}^{N} P_i = 1$. A classical example of order parameter in statistical mechanics is the magnetization M for spin systems such as an Ising model which gives account of the total orientation of the spins. Using the magnetization, the free energy of the system can be expressed as a function of it, so that the statistical mechanics of the disorder-order phase transition can be fully studied. Similar parameters are difficult to be conceived for proteins for the simple reason that proteins are heterogeneous objects whose order is not universal. However, quantities such as the number of native contacts the RMSD with respect to a reference native structure play the role of projection of the free energy landscape [Shea and Brooks, 2001] and mimic the role of order parameters. Typically free energy profiles are estimated from these projected quantity by computing the $-k_B T \ln$ of the histogram values. Because of their uni-dimensionality these target based quantities are able to distinguish an ordered phase (the target) from a "rest". The rest is the unfolded state which for its very nature is heterogeneous. Thus, if one is interested to a detailed description of the whole configurational space of proteins, projected quantities are insufficient and often ill defined. A partial solution to the lack of order parameters for folding is the use of *descriptors* based either on the clustering or the coarse-graining of the configurational space. Cluster analysis are usually based on structural similarity (RMSD, dRMSD). Coarse-graining methods basically consist in a *symbolization* of the configurational space. The symbolization of the configurational space corresponds

[1] The definition takes place from the assumption of extensivity of the entropy function in a canonical ensemble [Landau and Lifshitz, 1980].

[2] In a protein one usually call it "effective energy", which is the sum of the internal protein energy plus the solvation energy

to discretize a subset of the degrees of freedom with a consequent reduction of the system complexity. Given the continuous state space of a complex system, the idea is to partition it into a finite number of regions, each of which is labeled with some symbols. Symbols define discrete states that may or may not be directly related to the degrees of freedom of the system. These type of descriptions give arise to a "mesoscopic" or coarse-grained partitioned description of the configurational space. Mesoscopic means that probabilities are defined (or measured) out of all the possible outcomes of the descriptions. The multidimensionality of proteins entails that the descriptors can be related to the degrees of freedom in a non trivial manner. However, if the descriptor admits a finite set of states (as in symbolic states) the law of the total probability can be applied. According to the law of total probability given a set of finite or countable partitions B_i with $i = 1, 2, 3, ...$, if the events B_i can be measured, then for any event A it holds the relation $P(A) = \sum_i P(A|B_i)P(B_i)$ [Gnedenko, 1954]. Applying the theorem on a molecule whose configurational probability is given by the Gibbs-Boltzmann distribution, leads to the relation

$$\rho(\Gamma) = \sum_{i=1}^{N} \rho_i(\Gamma) P_i \tag{1.6}$$

where $\rho_i(\Gamma)$ is a conditional density within a partition i defined by the mesoscopic descriptor. The formula 1.6 is the key to interpret the link between the system and its observer in the context of thermodynamics. From that relation one is not only able to infer the thermodynamics as a direct consequence of the descriptors but also to judge the quality and the amount of information extracted from the system. In chapter 2 the consequences of this relation are investigated from the point of view of the thermodynamic analysis of folding simulations. In particular four kind of descriptors are studied (Cα-RMSD, Strings of native contacts, strings of secondary structure, strings of main chain rotational angles) on equilibrium molecular dynamics simulations of the GSGS peptide, the widely studied β-sheet mini protein [Ferrara and Caflisch, 2000, Rao and Caflisch, 2003, Rao and Caflisch, 2004, Rao et al., 2005]. Among the consequences of the law of total probability there is that each mesostate i has its own free energy, enthalpy and entropy. Approaches to the investigation of the protein thermodynamics have led to the development of graphical methods to analyze the topography of complex energy landscapes such as the disconnectivity graphs [Wales et al., 1998, Krivov and Karplus, 2002, Rylance et al., 2006] or the network representations [Rao and Caflisch, 2004, Gfeller et al., 2007]. These methods have revealed in a striking *visual* manner the heterogeneity of the configurational space of proteins and the multiple presence of different free energy basins shaping the energetic landscape. Without these studies such findings could only be guessed on the basis of solely one-dimesional descriptors such as the number of native contacts, the RMSD or the radius of gyration. The problem of reaction coordinates is closely connected to the understanding of folding kinetics and with the determination of hypothetical transition state ensemble TSE. In particular, if a TSE exists, quantities such as the folding probability P_{fold} can be largely useful for its structural characterization. A TSE is by definition a saddle point in a free energy profile so that a protein conformation standing on it would have half chances to fall into the folded state and half chances to fall into the unfolded. Methods to calculate P_{fold} from computer simulations have been proposed so far either from non-equilibrium parallel folding simulations [Du et al., 1998, Lenz et al., 2004] or directly from equilibrium folding simulations [Rao et al., 2005] (see chapter 3). The evident limitations of the P_{fold} approach is the implicit assumption of a two state folding so that the interpretation of 0.5 values, if the presence of intermediate states characterizes folding, is made difficult. On the basis of the experiments [Eaton et al.,

2000, Mayor et al., 2003] most of globular proteins are thought to have a two state kinetics, that is the overall kinetics governing folding is essentially single exponential meaning the a single main free energy barrier separates the unfolded phase from the folded. Deviations from this standard behavior are observed going from multi-exponential kinetics [Wagner and Kiefhaber, 1999, Sánchez and Kiefhaber, 2003] to non-exponential [Sabelko et al., 1999, Ihalainen et al., 2007]. How to reconcile these findings with the particularly rich free energy landscapes obtained from simulations? Attempts along this line are the models that use a Markovian approach to folding dynamics [Cieplak et al., 1998, Abundo et al., 2002, Swope et al., 2004a, Swope et al., 2004b, Park and Pande, 2006, Chodera et al., 2007, Noe et al., 2007]. The idea behind these models is to consider the configurational space of proteins discretized so that the dynamical transitions between two states at a certain time do not depend on the past history of the system. If the time is a continuos variable a master equation describes the system kinetics

$$\frac{d}{dt}P_i(t) = \sum_{j=1}^{N} (T_{ij}(\tau)P_j(t) - T_{ji}(\tau)P_i(t)) \tag{1.7}$$

where N states are introduced and $P_i(t)$ is the population of the ith state at time t, τ is a mesoscopic time scale and $T_{ij}(\tau)$ are transition probabilities. The transition matrix $T_{ij}(\tau)$ describes the mesoscopic dynamics of a single molecule in the configurational space. Typically, starting from equilibrium molecular dynamics simulations the configurational space is first clustered, or partitioned, and later the transition probabilities are estimated from the cluster time series. Ideally the equation 1.7 contains all the information and the pathways through which a protein molecule can reach the folded state. In practice, it is not trivial to find a partitioned representation of the configurational space which automatically leads to a Markovian dynamics (history independent), for the simple fact that in a simulation one cannot assume an overall diffusive regime due to finite size effects. When a protein is traveling in a free energy minima a diffusive regime and the Markov property can be assumed, because within a minima the protein can "forget" its previous trajectory. Conversely, when the protein is crossing a saddle area of the free energy landscape, to jump into the next stable state, the regime is ballistic and the trajectory is definitely non-Markovian [Plotkin and Wolynes, 1998]. A trivial example of diffusive regime is the equilibrium in which the system is relaxed in its stable states. In molecular dynamics simulation the situation is made worse by the fact that, basically the saddle points of a free energy landscape are very bad sampled. The main point is that a wrong hypothesis on the underling mesoscopic dynamics (non-Markovianity) leads to inconsistent solutions of the master equation or to factitious folding mechanisms. In this thesis the problem of the Markovianity from simulation is addressed and partially solved by developing a procedure that redefines the time series of the mesostates (either symbolic or from clustering) according with their "causal" relations. The procedure is called "causal grouping" and it allows to lump the mesostates that are non-Markovian in terms of the *future* they produce. The idea is extremely simple and allow to use the full potentialities of the master equation 1.7. From the master equation with N states one can easily calculate the spectral properties of the transition matrix which contains all the informations on the main kinetic modes of the system. Moreover, the mean first passage times MFPT for all the equilibrium transitions $i \rightarrow j$ can be easily calculated providing a full insight on folding kinetics. We applied this approach to all the proteins studied in this thesis. The approach has been also applied to coarse-grained simulations designed to study the aggregation of an amphipathic polypeptide. The coarse-grained polypeptides are characterized by a free energy profile having a distinct amyloid-

competent (i.e. beta-prone) state and an amyloid-protected state [Pellarin and Caflisch, 2006, Pellarin et al., 2007]. These simulations show that a decrease in the β-aggregation propensity results in a larger heterogeneity of elongation pathways, despite the essentially identical structure of the final fibril. In other words if the β-aggregation propensity is high the fibrils are formed through a simple deposition mechanism: the monomers first change state from aggregation-protected to β-prone and later rapidly polymerize until the mature fibril is formed; conversely if the β-aggregation propensity is low the system shows a lag phase during which monomers coordinate themselves into micelles and then, after the formation of critical nucleus, a proliferation of diverse pathways drive the system to the mature fibril. The application of the master equation on these simulations allowed to identify and quantify which leading mechanisms determine the elongation rate of the fibril formation (see chapter 6).

The application of the master equation or Markov chains represents an alternative way to address the problem of reaction coordinates. Many methods to find reaction coordinates that can be found in the literature are based on Bayesian optimization procedures or neural networks learning[Best and Hummer, 2005, Ma and Dinner, 2005, Peters et al., 2007]. In our opinion the master equation method, provided that the problem of Markovianity is kept under control, is indeed simpler, physically well established and gives account of any complex multi-exponential kinetics. After all we strongly believe to the Occam's razor reasoning, namely *It is vain to do with more what can be done with less*. In developing our causal grouping with the aim to construct a minimal Markov chain to describe the simulation data, we realized that the elaboration of a general theory called *computational mechanics*, proposed by Crutchfield and collaborators [Crutchfield and Young, 1989, Shalizi et al., 2002, Shalizi and Crutchfield, 2002], is ongoing. The theory proposes a technique that directly reconstructs minimal equations of motion from the recursive structure of measurement in a time series, in other words this method allows to discover hidden patterns in any kind of time series and translate them in a minimal Markov process. Let us briefly summarize the main points. Consider a discrete time and discrete-valued stochastic process in a time series $\cdots S_{-2}S_{-1}S_0S_1S_2\cdots$ with S_i a variable taking values in a finite alphabet of λ symbols. Given any time t the time series can be always divided in a past $\overleftarrow{S}_t$ and a future $\overrightarrow{S}_t$. Now, the conditionally stationarity of the process is assumed, which means that for any future event the conditional probability $P(\overrightarrow{S}_t \mid \overleftarrow{S}_t)$ does not depend on the time t. Given a certain past $\overleftarrow{S}$ one wants to make predictions of the future $\overrightarrow{S}$. The idea is to introduce a temporal partitioning of the time series in order to define some causal states as a basis of underling minimal Markov chain. Causal states are defined by first selecting a time window τ in which the differential conditional probability between two pasts is $|P(\overrightarrow{S}_\tau \mid \overleftarrow{S}'_\tau) - P(\overrightarrow{S}_\tau \mid \overleftarrow{S}''_\tau)| < \epsilon$ where ϵ is a sort of fitness parameter so that $\overleftarrow{S}'_\tau$ and $\overleftarrow{S}''_\tau$ belong to the same causal state. Essentially the causal states are a group of pieces of time series having length τ which lead to the same future $\overrightarrow{S}$. One can see that for any stationary process it can always be found a time window τ such that the obtained causal states generate a Markov process. The limitation of this method is to consider an overall τ time scale in defining the causal states. This can represent a delicate issue when multi-scaling time series are studied, meaning processes showing several time scales other than that necessary to the system to lose its memory. Folding trajectories are examples of multi-scale time series because of the complex structure of the free energy landscape: different minima or kinetic traps can in general posses quite different relaxation times. Recently a multi-scale version of the computational mechanics has been proposed to address this problem in the analysis of single molecule folding time se-

ries [Li et al., 2008]. A natural application of this approach are the single molecule florescence methods based on FRET (Foerster resonance energy transfer), ET (electron transfer) and fluorescence correlation spectroscopy [Talaga et al., 2000, Schuler et al., 2002, Lipman et al., 2003, Haran, 2003, Yang et al., 2003, Neuweiler and Sauer, 2004]. This new experimental technics represent important breakthroughs in folding investigation. Measured FRET efficiencies are proportional to the inverse 6th power distance between donor and acceptor fluorophore, thus the data measurements can be regarded as time series of intermolecular distances. Similarly the ET technic, measuring the phosphorescence lifetimes and single molecule electron transfer, is also capable to extract time series of distances. Despite the exciting possibility to observe a molecule in a non-ensemble context, these technics are not free from interpretation problems. An example of that is the degeneracy problem which means that corresponding to measured distances there can be ensembles of configurational states that fit within the same distance constraint. Another delicate aspect regarding the time series from single molecule experiments is that although all the measurements are at single molecule level, two time adjacent measurements might not be referring to the same molecule. In this respect, processing single molecule time series by means of learning algorithms if on one hand can better elucidate the diffusive character of the folding transition, on the other hand it does not necessarily lead to a detailed description of the system complex dynamics. The chances to construct more descriptive models from single molecule measurements would be indeed increase by the possibility to measure multiple intermolecular distances simultaneously [Schuler and Eaton, 2008].

1.4 Simplified protein sequences as a strategy to study folding

As mentioned above, proteins are the product of the biological evolution. Sequences evolved under the strong pressure that folding must be an efficient process to allow proteins exerting their biological function. Due to their multidimensional character, proteins have often been approached as a special version of spin glasses, thus as systems having a natural predisposition to be energetically frustrated [Bryngelson and Wolynes, 1987, Buchler and Goldstein, 1999a]. For globular proteins where a folded structure is a prerequisite for functioning, folding must be a "robust" process meaning that the amino acid sequences have been selected in such a way the frustration was minimal [Klimov and Thirumalai, 1996]. Robustness in this context means the necessity that potentially dangerous configurations (for instance those promoting aggregation) must be avoided during folding and that the native configuration has to be reached in a smooth way. In our opinion the evolution of protein sequences is not very alike from that of natural languages. Some researchers think that the evolution of natural languages might be the product of a universal tendency to a dynamic equilibrium between syntactic content (the punctuation) and the semantic content which is the meaning of the message that a speaker efficiently transmit to another speaker [Crofts, 2007, Lieberman et al., 2007]. In protein sequence evolution of the syntactic structures are those represented by the physicochemical properties of the sequence of amino acids (their hydrophobicity, steric hindrances, their propensity to assume local structures, etc.) while the semantic content represents the function proteins have evolved for. Likely the sequence evolution could have been characterized from an harmonization-competition of these two aspects: the structural features of proteins versus their purpose. In globular proteins the equivalence between structure and function suggests that the syntactic organization (the three dimensional protein structures) tightly correlates with

the semantic content (the protein function). On the other hand the existence of the intrinsically unstructured proteins suggests that the development of a complex semantic content (multiple functional role of these proteins) at the expense of a syntactic order (3D ordered structures) might be the final treat of an evolutionary path driven by the *meaning of the message*. As Shannon first understood [Shannon, 1948] communication requires two components, a thermodynamic framework for coding and transmission, and a semantic content which has to be recognized by a counterpart in a communication channel. Proteins are made up of sequences of amino acids and amino acids are chosen out from an alphabet of 20 types. If we focus on globular proteins in which the structure is completely encoded in the sequence then a question naturally arises: how many residue types are really necessary to encode the structural content? Protein sequences, as heritage of the genetic code, represent the words and the sentences of the complex language of living matter. As in any natural language with its syntactic rules, meaningless words can be constructed out of the combinations of letters as well as meaningless sentences can be built out of well formed words. On the other hand new words or neologisms can be included or not into the language inasmuch as they are accepted or not by the community that utilizes the language. Since in globular proteins the syntactic structure of sequences likely contains the message, its three-dimensional and functional structure, then a more specific question can be asked: what is the minimal number of residue types necessary to fold a protein? This question has been addressed both experimentally and theoretically so far, especially in the topic of protein design. Experimental studies conducted by Davidson and coworkers on random libraries of sequences with only three amino acids constitute a remarkable example of this research topic [Davidson and Sauer, 1994, Davidson et al., 1995, Cordes et al., 1996]. In these studies a library of synthetic genes encoding 80- to 100-residue composed mainly of random combinations of glutamine Q, leucine L, and arginine R were expressed in *Escherichia coli*. Among the proteins obtained some (about the 1% on a huge library) QLR proteins were well expressed and well characterized. These proteins, although totally artificial, have been shown to posses high helical content from CD measurements. These studies led Davidson and collaborators to conclude that the key elements of protein design seem to be the proper placement of hydrophobic residues along the polypeptide chain and the ability of these residues to form a well packed core. On another line are the works carried out in the group of Baker. In [Riddle et al., 1997] a β-sheet protein, the SH3 domain, was simplified by using 5 letter amino acids: Isoleucine I, Lysine K, Glutamic Acid E, Alanine A and Glycine G. The study was conducted using a phage-display selection strategy to promote the biological protein activity. The use of the residues I, K and E was justified by the fact that globular proteins contains non-polar interiors and polar exteriors so any experimental simplifying framework should contain both polar and non-polar residues. Alanine and Glycine were the better conserved residues in the combinatorial libraries. Despite the dramatic change in sequence, the folding rates of the simplified versions of the SH3 protein were very close to that of the wild type. NMR analysis showed a well packed core which justified the high protein stability. Thus the selection procedure eliminated molten globular structures in favor of function. Finally, Baker and coworkers argued that simplified sequences constitute an opportunity to investigate the evolution of the rapid and cooperative folding of small proteins. In order to exert their function globular proteins need a folded state both stable and kinetically accessible. While the former is under evolutionary pressure it is still unclear whether the latter is also an evolutionary factor. Baker and collaborators [Plaxco et al., 1998, Watters and Baker, 2004] also stressed that the number of letters required to obtain a foldable sequence could not be lowered to 3, they were

unable to obtain foldable sequence containing only one polar and two non-polar amino acids. A point this that is reinforced by Wolynes in [Wolynes, 1997] in light of the energy landscape theory: too simplified sequences would not have an enough "stability gap" , the energy difference between the native state and the rest of the configurational space, to guaranty thermodynamic control. Three letters alphabets have also benn excluded from lattice simulations in the protein design investigations conducted by Shakhnovich [Shakhnovich, 1998]. However in Wolynes's concluding remarks it is claimed that particularly symmetric structures could be encoded in a 3 letter amino acid alphabet (see for example the design a four helix bundle using only a 3 letter alphabet [Regan and De Grado, 1988]), a fact that would possibly bring into question the role of the hydrophobic code in protein folding. The studies of Rose and coworkers have often stressed on the hierarchical character of the folding process [Baldwin and Rose, 1999b, Baldwin and Rose, 1999a, Fitzkee et al., 2005, Rose et al., 2006]. In their view the unfolded state is pre-organized by the local propensities of short sequence stretches which drive the formation of local order. Typically the formation of local order drives the early coordination of the secondary structures which successively get assembled into the folded configuration hierarchically. According to Rose and coworkers the main driving force of the hierarchical assembly is the backbone hydrogen bonding which is responsible of pre-sculped configurational spaces. A consequence of this view is that, likely only a few number of amino acids, organized in "syntactically" well formed sequences, can give pre-sculped configurational spaces made of several minima, so that the role of side chains is that of selecting a specific minima. In a recent computational work it has been suggested that secondary structure propensities determines protein topologies [Fleming et al., 2006]. In that work the correct secondary structure assignments were used to constrain polypeptide backbone chains devoid of side chains, and the most favorable folded conformations were determined by using Monte Carlo simulation with energy terms depending on molecular compactness, steric exclusion and backbone hydrogen bonding. Interestingly a small number of energetically favorable topologies were found for a set of 13 proteins and, in the majority of the cases, the native one was prominent. Early works of Finkelstein and Ptitsys more than 20 years ago on the study of folding patterns already suggested that the limited set of folding structural pattern might not directly linked with the amino acid sequence detail [Finkelstein and Ptitsyn, 1987, Chothia and Finkelstein, 1990].

In this thesis a method is proposed to study the folding mechanisms of simplified protein sequences by means of molecular dynamics simulations in implicit solvent. It is known that with the current computers power folding simulations are limited to small system (see [Caflisch and Paci, 2004] and references therein). Even though modern force fields were accurate enough to encode the folded state of small proteins (about 60 residues), the folding rate of these proteins could be so low that the computational time to fold them might be prohibitive. Another aspect is the role played by the inaccuracies of the force field, underestimation and overestimation of the strength of specific non-bonded interactions may lead to the increasing of the free energy frustration of the modeled protein with a consequent dramatical decrease of the folding rate. Here we studied the folding mechanisms of four β-sheet and one α/β proteins with simplified sequences that adopt only three residue types. The choice of the residue types was done according with their secondary structure propensities, in particular sequences with three residue types were constructed from the list of alanine for helices, glycine and serine for turns and threonine for β. Full β proteins have respectively 20, 28, 36, 44 residues and were designed to fold in a three-, four-, five-, six-stranded β-sheets. The α/β protein has 56 residue and its sequence is a simplified version of

the B1 domain of protein G [Gallagher et al., 1994]. Strikingly all the simplified proteins reversibly fold in simulations, the full β proteins in a double-, three-, four-, five-stranded β-sheet respectively and the α/β in the native topology of the B1 domain of protein G. All the starting conformations of the folding simulations were extended. It will be shown that among the peculiarities of these proteins, the very low energetic frustration due to the simplified sequences leads to very high folding rates. In particular the folded states for these proteins resemble that of a molten globule, namely a native like secondary structure, the lack of specific tertiary contacts and the absence of a well defined hydrophobic core.

1.5 The structure of this thesis

Following the current introduction six other chapters compose the corpus of this thesis. In the second chapter the problem of the coarse-graining of the configurational space of polypeptides and its statistical mechanics are investigated in the context of molecular dynamics simulations. The problem of the description of complex systems is addressed with the help of information theory. Entropy based quantities are introduced to quantify the issues of the complexity and order of protein configurational space. A method based on Markov master equation is proposed to describe the complex dynamics of polypeptides. Such method is based of a causal definition of the configurational space which allows to to use the Markov hypothesis and thus providing a general framework to investigate protein folding kinetics. All the results in this chapter concerns the small 20 residue GSGS peptide widely studied in the Caflisch group. Based on this thesis chapter a paper is in preparation. In the third chapter a strategy is presented to study simplified proteins. The proteins are simplified in the sense that their sequences are constructed on an amino acid alphabet of solely three letters. Two kinds of proteins are studied: four modular full β-sheet proteins and a α/β protein which is the simplified version of the B1 domain of protein G. Molecular dynamics simulations of the simplified proteins are presented and the trajectories analyzed with the methods introduced in the previous chapter. This thesis chapter constitutes the basis of a submitted manuscript for publication that is included in chapter 4. Chapter five is a published article regarding the estimation of the P_{fold} for equilibrium MD simulations. In that article the present author contributed in developing the main article-idea and its theoretical support. The sixth chapter is a publication on the subject of pathways and kinetics of amyloid fibril formation through a coarse-grained peptide simulation model. This paper is the natural continuation of a previous one where Dr. Riccardo Pellarin has presented a mesoscopic model for the investigation of peptide aggregation. In that paper the present author has applied the ideas of the master equation as a descriptive tool to uncover the pathway complexity in the fibril formation. In the last chapter general conclusions and outlines are drown.

2 Protein folding mechanisms from MD simulations

2.1 Introduction: reducing the complexity

Understanding the mechanisms upon which a protein can reach its folded state is certainly one of the most important aims of computational biophysics. Though computer simulations give in principle access to the atomistic details of a protein conformation space they do not reduce its intrinsic complexity. Stuctural complexity arises from the number of degrees of freedom that give a full account of a protein structure. In a more physical language a specific conformation is nothing else than a "configurational microstate" (CM). Thus, for instance, if we take a protein having N atoms, then a CM is unequivocally determined by an euclidean vector whose elements are $3N - 6$ internal coordinates. If we consider the torsional angles of the chain molecule, then we have what can be called a rotameric microstate (RM), which is merely a vector of torsional angles whose length is $\sum_l^L R_l$ where R_l is the number of torsional angles of amino acid l. The latter representation of a RM is a common way to establish the statistical theory of chain molecules [Flory, 1969]. The full structural account is however not useful if one is interested to study average configurational properties to provide a statistical picture of the protein folding phenomena. This remark appears appropriate considering that our object of analysis are all atom molecular dynamics (MD) simulations in which all the degrees of freedom are taken into account. Thus, if one is interested in the statistical mechanics of proteins, the definition of some coarse grained states out of the ensemble of microstates sampled in a simulation is the crucial step to reduce the complexity of the configurational space of a macromolecule. Defining coarse grained states conceptually means to pass from a continuous description of the states to a discrete one in which some probability functions are defined according to some subjective physical "properties" [van Kampen, 1981]. In this chapter we investigate and define methods to classify the ensemble of microstates that are collected by means of a MD simulation. Moreover the descriptions introduced will make full use of the temporal causality of the simulation such that the kinetics of folding will be fully characterized. Within the general term of coarse grained descriptions we will distinguish between the "clustering" of the conformational space and discretization/symbolization of a subset of the degrees of freedom of the polypeptide. Both descriptions efficently classify the conformational space, being able to detect its main structural motifs. It will be clear along the chapter that both descriptions can be suitable in elucidating what are the main events that characterize the simulation under investigation.

In this chapter we focus on a 20-residue peptide designed by Rico and coworkers [De Alba et al., 1999] and referred below as GSGS (sequence **Thr1-Trp2-Ile3-Gln4-Asn5-Gly6-Ser7-Thr8-Lys9-Trp10-Tyr11-Gln12-Asn13-Gly14-Ser15-Thr16-Lys17-Ile18-Tyr19-Thr20**). Using MD simulation with a simple

implicit solvent based on the CHARMM force field [Ferrara et al., 2002], this peptide reversibly folds to a triple-stranded β sheet structure that fully agrees with experimental NOEs. The GSGS peptide has been extensively studied by MD simulation and its *in silico* folding and unfolding characteristics were previously described [Ferrara and Caflisch, 2000, Cavalli et al., 2002, Paci et al., 2003, Cavalli et al., 2003]. Such detailed study has been possible because MD trajectories including hundreds of folding events can be obtained within a few months of CPU time on a fast workstation.

With the model used here the peptide folds, in the sense that at the (melting) temperature of 330 K it populates a well defined conformational state, that we identify as the folded state, for about 50% of the simulation time. The rest of the time the peptide explores a large ensemble of conformations that can be quite different from the native state. Clearly, the notion of native and denatured state depends on the descriptors used to classify the conformations. The description of the configurations explored in the simulations, and the dependence of calculated folding rates on the descriptors, are among the topics of this chapter. The various methods of coarse grained descriptions used here, despite broad differences, all discriminate the native state as the one with lowest free energy, without using any reference to an experimental or *a priori* known structure.

2.2 Methods based on structural similarity

A widely used method to classify the structures explored in a simulation consists in comparing pairs of structures by computing the mutual positional root mean square deviation (RMSD) after least square overlapping them. Using the RMSD as a measure of similarity between structures, various *clustering* procedures have been proposed to pool structures that are mutually close, i.e., structurally similar (see e.g. [Daura et al., 1999b]). Clustering procedures assign the cluster *membership* to each structure, and in many cases they also define cluster centers or cluster representatives. A crucial issue consist in the definition of a cutoff, i.e., a threshold RMSD value below which two structures are considered similar. If a too small cutoff is chosen, the clustering does not achieve its primary objective of considerably reducing the number of distinct conformers, and it may produce an excessive number of clusters with a single member ("singletons"). If a large cutoff is chosen, many structure pairs with rather distinct properties, e.g. structures belonging to different low-energy basins, will be assigned to the same cluster. Since there is no *a priori* definition of a suitable, global RMSD cutoff, its choice is guided by efficiency considerations, and it represents a balance between computational speed and reduction of complexity on one side, and the detail or significance of the results that are sought on the other side. In Figure 2.1 the pairwise Cα-RMSD distributions of the conformations sampled in the trajectory of the GSGS are shown. The multiple peaks are a consequence of the fact that the conformation space is not uniformly populated but there are local minima or basins in the free energy surface. The first peak in the Cα-RMSD distributions corresponds to the typical Cα-RMSD distance between pairs of structures within one low-energy basin. Although the global distribution of pairwise Cα-RMSD reflects an average over all the low free energy basins, which might have different sizes, using a cutoff that is larger than the first minimum in the global RMSD distribution might improperly assign many structures of distinct free energy basins to the same cluster. Furthermore, the thermodynamic and kinetic properties of the system, as derived from the clustering data, depend on the cutoff radius used, and too large cutoffs

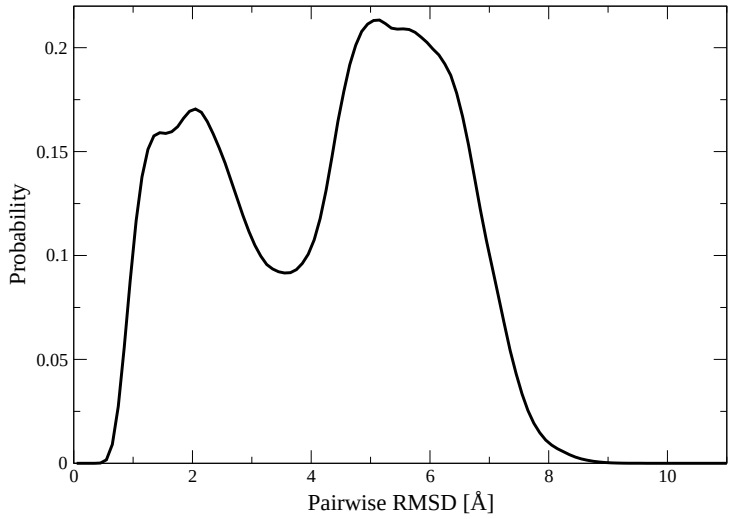

Figure 2.1: The distribution of the pairwise Cα-RMSD computed from an equilibrium simulation of the GSGS peptide.

could lead to inconsistent results. From the pairwise distributions in Figure 2.1 one can derive a critical cutoff encompassing most structure pairs within single basins of about 1-2 Å. Though sophisticated methods of clustering exist (see e.g. [Domany, 1999]), where a size for the coarse graining does not need to be defined beforehand, we chose to employ a more simplistic approach; it is more intuitive and good enough for our purposes.

On the technical viewpoint, the clustering of similar structures was performed using the program CLUSTER [Schaefer et al., 2006]. The method groups conformations based on their pairwise Cα-RMSD distance using a single, global cutoff [Daura et al., 1999a]. In this work, only the Cα atom positions were used to define the conformation and to calculate RMSDs. The program is able to cluster large numbers ($> 10^6$) of conformations from long-time simulations, e.g. a ~10 μs trajectory with coordinates saved every 20 ps.

Cluster centers are determined by an approach consisting of two nested iterations. In the first, "major" iteration level, a subset of conformers is determined by sampling the trajectory every Nth frame that has not yet been assigned to a cluster, e.g., every 10 or every 100 not-yet-clustered structure. The internal iterative procedure determines new cluster centers by evaluating the Cα-RMSD between all pairs of structures in the subset. Using RMSD$_{ij} < R_{cut}$ to cluster, the conformer i with the largest number of neighbors j in the subset is the first new cluster center. All frames whose distance from i is below the cutoff are then removed from the subset, and the next cluster center is again identified as the frame with the largest number of neighbors in the remaining subset. This step is repeated until only singletons remain in the subset, i.e., conformations that have no other structure within the cutoff distance [Daura et al., 1999a]. Subsequently, all conformers (including those outside the current subset) within cutoff distance from any of the new cluster centers are removed from the entire trajectory. The next major iteration begins by reducing the frequency of selecting unclustered conformers by a factor of 2. The algorithm is repeated until all structures have been assigned to a cluster, or until the selection frequency at the last major iteration is 1 such that all remaining frames have been considered as potential cluster

centers. For the cluster assignments the algorithm used here keeps only the identity of the cluster centers from the iterative procedure, and then assigns all frames of the trajectory to the nearest cluster center, i.e., the center with the minimum Cα-RMSD distance. This clustering approach has the advantage that the cluster assignment is not biased with respect to the order in which the cluster centers are identified. The assignment to the nearest cluster center is reminiscent of the calculation of Voronoi volumes for a set of points in the Cartesian space, where every volume element is also assigned to the nearest point [Voronoi, 1908]. This clustering program thus generate cluster representers and occupation probabilities that are well suited for a comparison between conformational substates and thermodynamic analysis. Currently, only the pairwise Cα-RMSD between conformations is implemented in CLUSTER as a measure of similarity. The program can be easily extended to use other measures, e.g. some measure of similarity of the secondary structure elements that are formed, the mean deviation of atom-pair distances from those in a reference structure to monitor domain motions, or the Euclidean distance of the peptide backbone ϕ/ψ angles [Karpen et al., 1993].

2.3 Symbolization of the conformation space

A limitation of clustering methods based on structure similarity is that the whole structure is usually compared and the local structural similarities do not emerge spontaneously. With that aim we introduce a discretization approach for the configuration space based on local properties. The approach will lead us to the symbolization of the trajectories of microstates into extended strings of symbols. Symbolization of the degrees of freedom and the consequent reduction of the system complexity is a common tool of investigation in the study of dynamical systems [Robinson and Clark, 1999, Badii and Politi, 1999]. Given the continuous state space of a complex system, the idea is to partition it into a finite number of regions, each of which is labeled with some symbol. Symbols are therefore discrete states that may or may not be related to the degrees of freedom of the system. The symbolic states are built up according to properties that the observer subjectively decides to be important to describe the configurational space of a polypeptide chain. Symbolization leads to a loss of information about the system due to the change of description in the space state, notably from continuous to discrete values. Moreover symbolic dynamics often exhibits stochastic properties even when the underlying continuous dynamics is deterministic. An example of that is the motion of a classical particle which moves under the effect of the temperature in a double well potential. If one discretizes the single degree of freedom, then a Markovian dynamics in the coarse grained space can be obtained. In the context of polypeptide chains it is convenient to define an "alphabet" of the symbolic states at either single or pair residue level that combined, provide a "string" description of the polypeptide discrete micro-states. Three different ways to construct symbolic states will be introduced below, in particular they will make use of the list of native contacts, the secondary structure and the main chain torsional angles.

2.3.1 Strings of native contacts

Native contacts are a common way to identify the degree of nativeness of a polypeptide with respect to its folded state. Here, instead of considering them as a scalar quantity, regardless of the local details, we treat them in a different manner. States of native contacts are defined as contacts from all the pairs of Cα

atoms that are distant less than 8 Å apart, and separated by more than three residues along the sequence. The total number of contacts in the native reference structure (as considered in [Ferrara and Caflisch, 2000]) of the GSGS peptide is 40; out of these 40 contacts 20 are between strands 1 and 2, and 20 between strands 2 and 3. Thus a symbolic state of the native contacts can be given by a binary string (string of native contacts, SNC[2]) of length 40. The completely native state is given by a string of 40 "1" while a completely unfolded state by 40 "0". The total number of strings is then given by $N^* \sim 2^{40} \sim 10^{12}$ states.

2.3.2 Strings of secondary structure

The secondary structure provides another way to construct symbolic states out of conformations which are supposed to be structurally similar. Using the DSSP [Kabsch and Sander, 1983] "alphabet" each residue of a protein can be either of eight symbols - (coil), `E` (extended strand in a β ladder), `S` (bend), `T` (hydrogen bonded turn), `B` (residue in isolated β-bridge), `G` (3_{10} helix), `H` (α helix), `I` (π helix), and each conformation identified by an octal string (string of secondary structure, SSS[8]). Assuming that two conformations correspond to the same state (when all the residues are in an identical secondary structural state) the maximum number of states for a polypeptidic chain of length 20 is $8^{18} \sim 10^{16}$ since the N-term and C-term residues have no assigned secondary structure as they have no ϕ and ψ respectively. The eight letters alphabet can be reduced to one based on four letters (SSS[4]) such as - (coil), beta (`E+B`), loop/turn (`S+T+G`) and helix (`H+I`). The symbol G is considered as turn-like configuration. In such a grouping scheme the total number of strings turns out to be $N^* = 4^{18} \sim 10^{10}$.

2.3.3 Strings of rotational states

An other way to symbolize the configuration space of a polypeptide is based on the discretization of the Ramachandran plot [Ramachandran et al., 1963] of the amino acids. The Ramachandran plot is a two-dimensional map representing the preferences of main-chain rotational angles (ϕ, ψ). Due to the constraints imposed by the steric hindrances in the backbone torsions, and by the local side chain-backbone interactions, only certain regions of a Ramachandran plot are allowed. Thus, we partition the Ramachandran plot corresponding to the most probable regions. An ensemble Ramachandran plot has been considered to choose the partitions. We computed the (ϕ, ψ) propensities for all the residues adding up their frequencies onto a 360×360 uniform grid. In Figure 2.2 is shown the ensemble Ramachandran plot with the estimated normalized frequencies $P(\phi, \psi)$; $P(\phi, \psi)$ is the probability of the dihedral angle pair (ϕ, ψ) estimated by counting the frequencies of any ϕ, ψ pair in the simulation. According to the number of frequency peaks, four regions can be distinguished on the ensemble Ramachandran plot which correspond to discrete rotational states, indicated as `0,1,2,3`. The barriers delimiting these regions are drawn according with the position of the regions of low frequency on the plot. The maximum number of states, or (strings of rotational angles, SRA[4]) for a 20-residue polypeptide is $N^* \sim 4^{18} \sim 10^{10}$ as the first and last residues have no ϕ and ψ angles respectively. The symbolization method is reminiscent of the Flory's rotational isomeric approximation [Flory, 1969, Flory, 1974] that was introduced to develop the random-coil statistical theory for chain molecules.

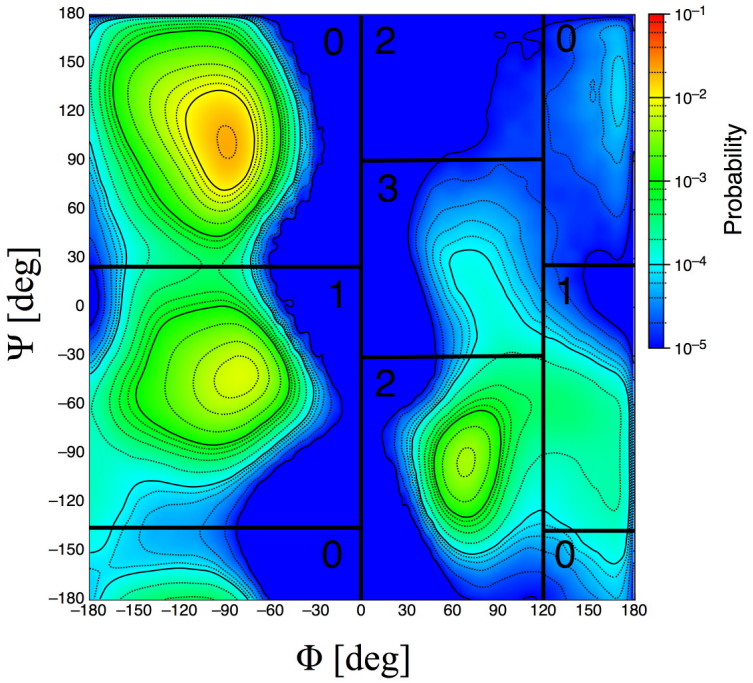

Figure 2.2: Ensemble Ramachandran 2D histogram with the partitions used to construct the strings of rotational states for a polypeptide chain SRA[4]. The 2D histogram is obtained from the cumulative frequencies of all (ϕ, ψ) pairs on a 360×360 grid. Frequencies are estimated from the GSGS MD trajectory.

2.4 Thermodynamics of the coarse graining

A coarse grained description automatically yields finite and countable the accessible states of the system, and defines a normalized probability function on them. On that respect, the new states cannot be considered as proper macrostates and the correct word to call them is *mesostates*, which means states defined by probabilities. Mesostates are discrete states and finite in contrast with microstates which are by definition continuous and infinite. Thus the coarse grained descriptions not only aim to reduce the complexity of the system, but also to use statistical mechanics to infer the thermodynamics. Yet, the problem is to define a theoretical support that allows to quantify and to understand the physical meaning of a coarse graining procedure, which are also strongly influenced by the observer.

2.4.1 Entropy, information and order parameters

What is the meaning of the coarse graining from a thermodynamic point of view? Let us consider an arbitrary complex physical system at its thermal equilibrium in the context of a canonical statistical ensemble. The Gibbs-Boltzmann probability density reads

$$\rho(\Gamma) = \frac{1}{Z} e^{-\beta E(\Gamma)} \tag{2.1}$$

which gives the amplitude of a particular microstate $\Gamma = (\mathbf{p}, \mathbf{q})$ in the phase space, E is the energy function of the system, $\beta = 1/k_B T$ and Z is the partition sum

$$Z = \int e^{-\beta E(\Gamma)} d\Gamma \tag{2.2}$$

The total amount of energy that is not available to perform active work is given by the thermodynamic entropy, namely by the celebrated formula

$$S = k_B \int_{\Omega} \rho(\Gamma) \ln \rho(\Gamma) d\Gamma \tag{2.3}$$

where k_B provides the units to measure it [kcal K^{-1} mol^{-1}] and Ω is the phase space. The total entropy defines the free energy of the system (namely the amount of energy available to work) such that

$$G = \overline{E} - TS \tag{2.4}$$

where $\overline{E}$ is the mean internal energy of the system [Landau and Lifshitz, 1980]. Moreover, from statistical mechanics one has

$$G = -k_B T \ln Z \tag{2.5}$$

In a coarse grained description the observables can be related to an experimental equipment and thus are able to classify the microstates in terms of their physical properties. The observable may be thought as an "order parameter" that we shall assume to be a discrete function on the phase space. However, an observable is always chosen by an *observer* who asks himself what kind of information can he gather from the system through the experiment. That is all but a mere philosophical issue, as all the results one can get from an experiment are strongly related to the initial hypothesis and the questions the observer assumed and posed himself. Since judging an observer is evidently outside the purposes of this work, we can try to analyze what are the consequences of certain assumptions made on observables or order parameters.

An example of order parameter is for instance the total magnetization in a 2 dimensional Ising model whose values depends on the local arrangement of the spins. Such a quantity is able to distinguish between two thermodynamic phases: the orientated and the disordered phases. An order parameter, hereafter called ξ, by definition maps the phase space Ω in partitioned subspaces ω_i such that a conditional density function may be defined according to the law of the total probability [Gnedenko, 1954]:

$$\rho(\Gamma) = \sum_{i=1}^{N} P(\xi_i) \rho(\Gamma|\xi_i) \tag{2.6}$$

where $P(\xi_i)$ is the normalized distribution of the order parameter values in the partitioned space $\Omega = \oplus_{i=1}^{N} \omega_i$ and $\rho(\Gamma|\xi_i)$ is the conditional density such that $\rho(\Gamma|\xi_i) \neq 0$ if $\Gamma \in \omega_i$ and identically zero otherwise. The subspaces ω_i and the distribution $P(\xi_i)$ define the mesostates of the coarse graining previously defined. Combining the equations 2.6 and 2.3, the total entropy splits in two parts

$$S = H + S^{\mathrm{b}} \tag{2.7}$$

where

$$\begin{aligned} H &= -k_B \sum_{i=1}^{N} P(\xi_i) \ln P(\xi_i) \\ &= \sum_{i=1}^{N} P(\xi_i) H_i \end{aligned} \tag{2.8}$$

is the Shannon entropy [Shannon, 1948] of the order parameter distribution, with $H_i = -k_B \ln P(\xi_i)$ the Shannon entropy of a sigle mesostate, while

$$\begin{aligned} S^{\mathrm{b}} &= -k_B \sum_{i=1}^{N} P(\xi_i) \int_{\omega_i} \rho(\Gamma|\xi_i) \ln \rho(\Gamma|\xi_i) \\ &= \sum_{i=1}^{N} P(\xi_i) S_i^{\mathrm{b}} \end{aligned} \tag{2.9}$$

is the sum of all vibrational entropies within the space partitions created by the order parameter ξ, that are

$$S_i^{\mathrm{b}} = -k_B \int_{\omega_i} \rho(\Gamma|\xi_i) \ln \rho(\Gamma|\xi_i) \tag{2.10}$$

Following the argumentations of information theory the entropy H of a certain variable measures the amount of information carried by the variable itself, or in other words its statistical uncertainty. Thus, large values of H means high uncertainty (and conversely high disorder, high information) while low values means that only few states of ξ are very populated and thus the uncertainty is low (high order and low information) [Khinchin, 1957]. In a more biophysical language the entropy H represents nothing else than an estimation of the configurational entropy of the polypeptide, as correctly pointed out by Karplus in [Karplus et al., 1987]. With configuration here we intend mainly backbone configurations, namely something *recognizable* within the cell. We shall note that Shannon entropy is upper bounded at its maximal value when all the mesostates are equally populated, $P(\xi_i) = 1/N$ In that case one has

$$H_{\max} = k_B \ln N \tag{2.11}$$

That maximal Shannon entropy $H_{\max}$ can be assumed as the conformational entropy of the random-coil state for the polypeptide, as by definition in the random-coil state of a polypeptide all the conformations are equally accessible. In the case of the vibrational entropy S_i^{b} the idea is similar: high entropy values means high disorder within the phase subspace ω_i. According to the second principle the transfer between the conformational entropy of the order parameter and the vibrational entropies is allowed [Ebeling, 1993]. The order parameter could be introduced naturally from the features of the system, for instance the magnetization vector in a two dimensional Ising model, or the orientational degrees in liquid crystals. In a more complex system such as a protein the arbitrary choice of the order parameter is both crucial and delicate for the kind of description one is interested in. That is also known as the problem of the right reaction coordinate in protein folding kinetics, which is also related to what a coarse graining should pick in a polypeptide.

We want to find an expression for the free energy which raises from the choice of a coarse graining. Let's consider first the internal energy of the system. The internal energy of the system is normally expressed as an average on the ensenble

$$\overline{E} = \int_{\Omega} E(\Gamma) \rho(\Gamma) d\Gamma \tag{2.12}$$

Replacing the equation 2.6 into 2.12 we obtain

$$\begin{aligned} \overline{E} &= \sum_{i=1}^{N} P(\xi_i) \int_{\omega_i} E(\Gamma)\rho(\Gamma|\xi_i)d\Gamma \\ &= \sum_{i=1}^{N} P(\xi_i)\overline{E_i} \end{aligned} \tag{2.13}$$

which means that the total internal energy is the weighted average of the internal mean energies within the subspaces that correspond to the order parameter. For the energy fluctuations we have

$$\begin{aligned} \sigma^2(E) &= \int_{\Omega} (E(\Gamma) - \overline{E})^2\rho(\Gamma)d\Gamma \\ &= \sum_{i=1}^{N} P(\xi_i) \int_{\omega_i} (E(\Gamma) - \overline{E})^2\rho(\Gamma|\xi_i)d\Gamma \\ &= \sum_{i=1}^{N} P(\xi_i)(\overline{E_i} - \overline{E})^2 + \sum_{i=1}^{N} P(\xi_i)\sigma^2(E_i) \end{aligned} \tag{2.14}$$

where $\sigma^2(E_i) = \overline{E_i^2} - \overline{E_i}^2$. Thus the order parameter splits the energy fluctuations in two parts, those internal to the partitions ω_i (*microscopic*) and those proportional to the deviation between the squared mean energies $\overline{E_i}$ and the squared mean total energy $\overline{E}$ (*mesoscopic*). The total free energy is then given by

$$\begin{aligned} G &= \overline{E} - TS \\ &= \sum_{i=1}^{N} P(\xi_i) \left(\overline{E_i} - TS_i^{\mathrm{b}}\right) - TH \\ &= \sum_{i=1}^{N} P(\xi_i)G_i - TH \end{aligned} \tag{2.15}$$

where we have defined the free energies of the mesostate

$$G_i = \overline{E_i} - TS_i^{\mathrm{b}} \tag{2.16}$$

within the phase partitions defined by the order parameter. Knowing that $G = -k_B T \ln Z$ then from equation 2.15 one easily obtains an expression for the total partition function

$$Z = e^{H/k_B} \prod_{i=1}^{N} Z_i^{P(\xi_i)} \tag{2.17}$$

with

$$Z_i = e^{-\beta(\overline{E}_i - TS_i^{\mathrm{b}})} = e^{-\beta G_i} \tag{2.18}$$

When no order parameter is introduced, then $H = 0$, $S = S_b$ and $Z = e^{-\beta(\overline{E} - TS)}$. For the probability distribution we can take

$$\begin{aligned} P(\xi_i) &\sim e^{-H/k_B} e^{-\beta(\Delta E_i - T\Delta S_i^{\mathrm{b}})} \\ &= \frac{Z_i}{Z} \end{aligned} \tag{2.19}$$

The relation 2.19 is well formed, in fact by taking the $\ln$ of both sides we obtain

$$\begin{aligned} T\Delta H_i &= \Delta G_i \\ &= \Delta E_i - T\Delta S_i^{\mathrm{b}} \end{aligned} \tag{2.20}$$

where $\Delta H_i = H_i - H$ is a conformational entropy loss/gain (negative/positive) from a decoy of mesostates to the ith mesostate. Depending on the sign of the conformational entropy this is compensated by a microscopic energy loss/gain $\Delta E_i = \overline{E_i} - \overline{E}$ and an internal entropy gain/loss $\Delta S_i^{\mathrm{b}} = S_i^{\mathrm{b}} - S^{\mathrm{b}}$ with respect the whole ensemble of microstates.

The entropy S and the free energy G can be estimated from the fact that the ensemble distribution of the energy follows a gaussian. The distribution of the total energy of the GSGS (which is the sum of potential and solvation energies) on the ensemble of microstates is a perfect gaussian with mean $\overline{E} = -4.04$ kcal/cal and standard deviation $\sigma(E) = 10.74$ kcal/mol (see figure 2.3). This fact is merely a consequence of the center limit theorem applied to the energy function, which states that any sum of variables however distributed defines a random variable normally distributed [Gnedenko, 1954]. The distribution in figure 2.3, hereafter called $k(E)$, represents thus the density of microstates of energy E in a canonical ensemble. Note that the partition function Z can be also written in terms of an integral in the space of the energies instead of one in the phase space such that

$$Z = \int_{-\infty}^{+\infty} k(E) e^{-\beta E} dE \tag{2.21}$$

[Huang, 1987]. In particular we take

$$k(E) = \frac{1}{\sqrt{2\pi}\sigma(E)} e^{-(E-\overline{E})^2/2\sigma^2(E)} \tag{2.22}$$

which is the density of microstates having energy E. Replacing the equation 2.44 into 2.21 and integrating we obtain

$$\begin{aligned} Z &= \frac{1}{\sqrt{2\pi}\sigma(E)} \int_{-\infty}^{\infty} e^{-(E-\overline{E})^2/2\sigma^2(E)} e^{-\beta E} dE \\ &= e^{-\beta\overline{E} + \frac{1}{2}\beta^2\sigma^2(E)} \end{aligned} \tag{2.23}$$

which compared with the relations 2.4 and 2.5 gives a total entropy

$$S = \frac{1}{2} k_B \beta^2 \sigma^2(E) \tag{2.24}$$

where $k_B\beta^2\sigma^2(E)$ is the specific heat C_V, that in the case of the GSGS at the working temperature $T = 330$ K is 0.52 kcal mol^{-1} K^{-1}. The equation 2.24 states that the total entropy is proportional to the fluctuations of the energy[1], in particular to the square of the ratio $\sigma(E)/k_BT$, which can be assumed as as frustration index of the energetic landscape, plus an unknown constant S_0 that we set to 0 for simplicity so that $S = 0.26$ kcal mol^{-1} K^{-1}. The word "frustrated" is here used as a counterpart of "disordered", i.e. highly entropic. We can now estimate also a total free energy from $G = \overline{E} - TS$ that is -91.4 kcal/mol. From the equation of the energy fluctuations 2.14 we obtain the bound entropies S_i^{b} within the mesostates (and furthermore $S^{\mathrm{b}} = \sum_i P_i S_i^{\mathrm{b}}$)

$$S_i^{\mathrm{b}} = \frac{1}{2} k_B \beta^2 \sigma^2(E_i) \tag{2.25}$$

where we have taken a gaussian density of microstates within a mesostate ω_i. An approximated relation between the Shannon entropy of the ensemble of mesostates and the *mesoscopic* energy fluctuations is

[1] From 2.21 one has $Z = \int k(E) e^{-\beta E} dE = \int e^{-\beta E + \ln k(E)} dE = \int e^{\beta(TS(E) - E)} dE = e^{\beta(TS(\overline{E}) - \overline{E})}$

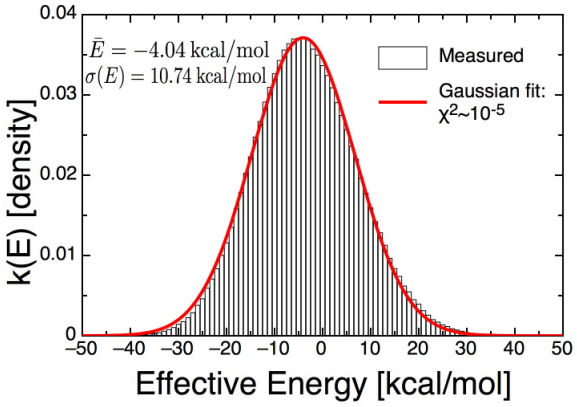

Figure 2.3: The distribution of the total energy (the sum of potential and solvation energies) from the whole ensemble of microstates of the GSGS simulation. The agreement with a gaussian is excellent, with mean effective energy $\overline{E} = -4.04$ kcal/mol and standard deviation $\sigma(E) = 10.74$ kcal/mol.

obtained

$$-k_B \sum_{i=1}^{N} P(\xi_i) \ln P(\xi_i) \sim \frac{1}{2} k_B \beta^2 \sum_{i=1}^{N} P(\xi_i)(\overline{E_i} - \overline{E})^2 \tag{2.26}$$

which implies

$$H_i \sim \frac{1}{2} \beta^2 (\overline{E_i} - \overline{E})^2 \tag{2.27}$$

The equation 2.27 means that mesostates having mean energy far from the total mean energy have high information content. Finally we call S^{m} the mesoscopic entropy

$$\begin{aligned} S^{\mathrm{m}} &= \frac{1}{2} k_B \beta^2 \sum_{i=1}^{N} P(\xi_i)(\overline{E_i} - \overline{E})^2 \\ &= \sum_{i=1}^{N} P(\xi_i) S_i^{\mathrm{m}} \end{aligned} \tag{2.28}$$

The mesoscopic entropy S^{m} can be interpreted as a mixing entropy of the ensemble of mesostates with well defined internal energies. Thus we have

$$S = H + S^{\mathrm{b}} = S^{\mathrm{m}} + S^{\mathrm{b}} + S_0' \tag{2.29}$$

with S_0 an offset. One can also see that $S^{\mathrm{m}} \lesssim S^{\mathrm{b}}$ for any random chosen $N \ll M$ mesostates.

When a coarse graining is introduced, a map of partitions is defined according to the order parameter ξ over the configurations space and the compute of the free energies computation should include both energy and entropy contributions as here shown. Quite usually in the literature the free energies are estimated from simulations by considering the H_i terms, which is in principle incomplete. The order parameter is a quite general quantity which can be either unidimensional or multidimensional. In case of the coarse grained descriptions of the configuration space of the GSGS peptide, the string based mesostates, we have a multidimensional descriptor whose interpretation in structural sense is straightforward. For a clustering the order parameter has a more delicate interpretation though it can

i	ΔE_i $^{(*)}$	$\sigma(E_i)$ $^{(*)}$	P_i	$T\Delta H_i$ $^{(*)}$	$T\Delta S_i^{\mathrm{b}}$ $^{(*)}$	ΔG_i $^{(*)}$	Mesostate
1	-4.9	8.8	0.014	-3.4	14.9	-19.9	01110111111111111111111111111011111111111
2	-5.1	8.9	0.011	-3.3	16.6	-21.8	01111111111111111111001111011111111111
3	-4.4	8.8	0.009	-3.1	14.2	-18.7	01110111111111111111011111011111111111
4	-4.4	8.7	0.008	-3.1	13.4	-17.9	01110111111111111111001111011111111111
5	-5.4	8.9	0.008	-3.0	15.9	-21.4	01111111111111111111011111011111111111
6	-5.9	8.8	0.007	-3.0	15.0	-20.9	01111111111111111111111111011111111111
7	-4.7	8.6	0.006	-2.9	12.3	-17.1	11110111111111111111111111011111111111
8	-4.6	8.8	0.006	-2.9	14.4	-19.2	01110111111111111110101111011111111111
9	-5.0	8.7	0.006	-2.9	13.2	-18.3	01110111111111111111111111111111111111
10	-4.9	9.0	0.006	-2.8	16.8	-21.9	01110111111111111101111111011111111111
11	-4.7	8.9	0.005	-2.8	16.0	-20.7	01111111111111111110001111011111111111
12	-5.3	9.0	0.004	-2.7	17.3	-22.7	01111111111111111101011111011111111111
13	-3.9	8.8	0.004	-2.7	14.1	-18.1	01110111111111111110001111011111111111
14	-6.1	9.2	0.003	-2.5	19.6	-25.9	01111111111111111110010111011111111111
15	-3.3	8.6	0.003	-2.5	12.1	-15.5	01110111111111111110011111011111111111
16	-2.2	9.3	0.003	-2.5	21.8	-24.1	00000000000001110100001111101111111111
17	-5.0	8.9	0.003	-2.5	16.2	-21.2	11110111111111111101111111011111111111
18	-6.0	8.7	0.003	-2.5	13.4	-19.4	01111111111111111111111111111111111111
19	-4.0	8.7	0.003	-2.4	13.3	-17.4	11110111111111111111001111011111111111
20	-4.7	8.6	0.003	-2.4	11.9	-16.7	11110111111111111111111111111111111111
21	-4.4	8.7	0.003	-2.4	13.8	-18.3	11110111111111111110111111011111111111
22	-4.7	8.9	0.002	-2.4	15.8	-20.6	01110111111111111101111111111111111111
23	-4.7	8.8	0.002	-2.3	14.6	-19.4	11111111111111111111001111011111111111
24	-4.6	9.1	0.002	-2.3	18.6	-23.3	01111111111111111101001111011111111111
25	-4.3	8.9	0.002	-2.3	15.9	-20.3	01110111111101111111111111011111111111
26	-2.9	8.5	0.002	-2.3	11.3	-14.2	01100111111111111111111111011111111111
27	-6.5	9.2	0.002	-2.3	19.7	-26.3	00000000000000000110101100110111111111
28	-3.7	8.8	0.002	-2.3	14.2	-18.0	01110111111111111100011111011111111111
29	-3.7	8.8	0.002	-2.3	14.5	-18.3	01110111111111111101111111011111111111
30	-4.4	8.9	0.002	-2.3	16.1	-20.6	01111111111111111100111111011111111111
31	-5.1	9.0	0.002	-2.3	17.2	-22.4	01111111111111111101111111011111111111
32	-2.2	8.7	0.002	-2.3	13.6	-16.0	01111001111110011111111111111011111111111
33	-5.8	8.9	0.002	-2.2	16.3	-22.2	11111111111111111111111111111011111111111
34	-4.4	8.7	0.002	-2.2	12.9	-17.4	11110111111111111111010111111011111111111
35	-3.8	8.9	0.002	-2.2	16.2	-20.1	01110111111111111111111111111111111111110
36	-4.5	8.9	0.002	-2.2	16.3	-20.9	11110111111111111111000111111011111111111
37	-3.2	9.1	0.002	-2.2	18.7	-22.0	01110111111111111111111111111011111111110
38	-5.9	9.2	0.002	-2.2	20.3	-26.3	01110111111111111111110010111011111111111
39	-4.9	9.0	0.002	-2.1	16.7	-21.7	01111111111111111111101111111111111111111
40	-3.5	8.9	0.002	-2.1	16.5	-20.1	01111111111111111111000111011111111111
41	-5.1	8.8	0.001	-2.1	14.8	-20.1	11111111111111111111011111011111111111
42	-2.8	8.7	0.001	-2.1	13.1	-16.0	01111011111100111111111111011111111111
43	-3.3	8.8	0.001	-2.1	15.1	-18.5	01110111111111111101001111011111111111
44	-4.0	9.1	0.001	-2.1	18.6	-22.7	01110111111111111111011111111111111111
45	-7.5	9.2	0.001	-2.1	19.7	-27.3	00000000000000000110101101101111111111
46	-2.9	9.1	0.001	-2.1	19.2	-22.2	00000000000001110100001110110111111111
47	-4.5	8.5	0.001	-2.0	10.8	-15.4	11111111111111111110001111011111111111
48	-5.3	8.9	0.001	-2.0	16.3	-21.7	01111111111101111111111111011111111111
49	-5.5	8.8	0.001	-2.0	15.0	-20.6	11110111111111111101111111111111111111
50	-5.0	9.1	0.001	-2.0	18.3	-23.4	01111111111111111000111111011111111111

Table 2.1: $^{(*)}$ [kcal/mol]. Thermodynamic parameters: $\Delta E_i = \overline{E_i} - \overline{E}$, $T\Delta H_i = T(H_i - H)$, $T\Delta S_i^b = T(S_i^{\mathrm{b}} - S^{\mathrm{b}})$ (see equation 2.9 for the definition of S^{b}), $\Delta G_i = \Delta E_i - T\Delta S_i^{\mathrm{b}}$. The list of the first 50 most populated mesostates of the GSGS for the coarse graining based on the strings of native contacts SNC[2] with all the thermodynamic quantities estimated according to the equations in the text.

be assumed to be a scalar quantity.

The definitions of mesostates given previously allow us to map the trajectories of microstates into time series of mesostates. Indicating $\mathbf{q}(t)$ a microstate at time t and ω_i a mesostate, one defines the counter function

$$\theta_i(t) = \begin{cases} 1 \text{ if } \mathbf{q}(t) \in \omega_i \\ 0 \text{ if } \mathbf{q}(t) \notin \omega_i \end{cases} \tag{2.30}$$

where $\omega_i = \omega_1, ..., \omega_N$, and N the total number of mesostates visited by the peptide along the trajectory.

i	ΔE_i $^{(*)}$	$\sigma(E_i)$ $^{(*)}$	P_i	$T\Delta H_i$ $^{(*)}$	$T\Delta S_i^{\mathrm{b}}$ $^{(*)}$	ΔG_i $^{(*)}$	Mesostate
1	-3.5	8.96	0.0745	-6.3	12.63	-16.15	-EEEESSEEEEEESSEEEE-
2	-4.6	8.94	0.0630	-6.2	12.48	-17.08	-EEE-STTEEEEESSEEEE-
3	-4.3	8.97	0.0364	-5.9	12.83	-17.16	-EEEESSEEEEE-STTEEE-
4	-5.2	9.04	0.0280	-5.7	13.76	-18.94	-EEE-STTEEEE-STTEEE-
5	-2.6	9.02	0.0247	-5.6	13.48	-16.10	-EEEESSEEEEEESSEEE--
6	-3.7	8.93	0.0189	-5.4	12.32	-16.04	-EEESSTTEEEEESSEEEE-
7	-4.4	9.04	0.0170	-5.4	13.81	-18.22	-EEE-TTTEEEEESSEEEE-
8	-3.5	8.99	0.0127	-5.2	13.13	-16.60	-EEEESSEEEE-STTEE--
9	-3.2	8.95	0.0126	-5.2	12.55	-15.79	-EEEESSEEEEESSTTEEE-
10	-3.1	9.06	0.0107	-5.0	14.05	-17.14	-EEE-STTEEEEESSEEE--
11	-4.0	9.02	0.0101	-5.0	13.45	-17.41	-EEE-STTEEEESSTTEEE-
12	-4.6	9.01	0.0093	-5.0	13.41	-18.01	-EEESSTTEEEE-STTEEE-
13	-3.9	8.95	0.0071	-4.8	12.56	-16.42	-EEEESSEEEEE-TTTEEE-
14	0.3	9.11	0.0069	-4.8	14.80	-14.48	--EEESSEEEEEESSEEEE-
15	-1.8	9.14	0.0068	-4.7	15.11	-16.95	--EEESSSEEEEESSEEEE-
16	-4.6	9.34	0.0067	-4.7	17.93	-22.49	-EEE-TTTEEEE-STTEEE-
17	-4.6	9.14	0.0066	-4.7	15.18	-19.76	-EEEESTTEEEEESSEEEE-
18	-0.5	9.31	0.0060	-4.7	17.60	-18.06	--EE-STTEEEEESSEEEE-
19	-4.0	9.18	0.0054	-4.6	15.73	-19.75	-EEE-STTEEEE-STTEE--
20	-4.1	9.01	0.0049	-4.5	13.41	-17.55	-EEE-STTEEEE-TTTEEE-
21	-2.6	9.14	0.0047	-4.5	15.21	-17.84	-EEEESSSEEEEESSEEEE-
22	-2.5	9.13	0.0047	-4.5	15.06	-17.58	-EEESSTTEEEEESSEEE--
23	-0.7	9.13	0.0042	-4.4	15.00	-15.75	-EEEESSEEEEEESSEE---
24	-0.4	9.12	0.0034	-4.3	14.86	-15.27	--EEESSEEEEE-STTEEE-
25	-2.3	8.91	0.0033	-4.3	12.01	-14.27	-EEEESSEEEEESSTTEE--
26	-3.7	8.97	0.0030	-4.2	12.82	-16.54	-EEESSTTEEEESSTTEEE-
27	-1.9	8.90	0.0030	-4.2	11.83	-13.74	--EEESSSEEEE-STTEEE-
28	-2.9	9.05	0.0029	-4.2	13.98	-16.85	----STT---EEESSEEEE-
29	-1.0	8.84	0.0028	-4.2	11.01	-12.02	-EEE-STTEEEESSSEEEE-
30	-0.8	9.33	0.0027	-4.1	17.76	-18.53	--EE-STTEEEE-STTEEE-
31	-3.7	9.03	0.0026	-4.1	13.70	-17.43	-EEE-TTTEEEESSTTEEE-
32	0.6	8.80	0.0026	-4.1	10.51	-9.90	-EEEESSEEEEESSSEEEE-
33	-3.6	9.08	0.0025	-4.1	14.31	-17.90	-EEESSTTEEEE-STTEE--
34	-1.2	9.35	0.0025	-4.1	18.11	-19.32	--EE-TTTEEEEESSEEEE-
35	-0.5	9.26	0.0024	-4.1	16.86	-17.39	-EEEESSEEEEESSSEEE--
36	-1.0	9.18	0.0024	-4.1	15.73	-16.71	--EEESTTEEEEESSEEEE-
37	-4.1	8.96	0.0023	-4.0	12.71	-16.80	-EEEESTTEEEE-STTEEE-
38	-2.4	9.18	0.0023	-4.0	15.69	-18.12	-EEEESSEEEEE-TTTEE--
39	-2.4	9.53	0.0022	-4.0	20.67	-23.12	-EEE-TTTEEEEESSEEE--
40	-1.8	9.44	0.0022	-4.0	19.33	-21.12	-EEE-STTEEEEESSEE---
41	-1.7	9.41	0.0020	-3.9	19.01	-20.66	-EEE-STTEEEESSSEEE--
42	-2.4	8.67	0.0020	-3.9	8.87	-11.29	-EEEETTEEEEEESSEEEE-
43	-2.4	9.34	0.0019	-3.9	17.95	-20.34	-EEEESSSEEEE-STTEEE-
44	-3.5	8.72	0.0019	-3.9	9.46	-12.95	-EEESTTTEEEEESSEEEE-
45	-4.1	9.22	0.0017	-3.8	16.24	-20.33	-EEESSTTEEEE-TTTEEE-
46	-1.9	9.05	0.0017	-3.8	13.92	-15.77	----SSS---EEESSEEEE-
47	0.0	8.82	0.0016	-3.8	10.77	-10.73	--EESSTTEEEEESSEEEE-
48	-0.7	9.21	0.0016	-3.8	16.11	-16.79	-EEEESSEEEEEESSEETT-
49	1.6	9.08	0.0015	-3.7	14.34	-12.78	--EEESSEEEEEESSEEE--
50	-1.2	9.31	0.0015	-3.7	17.59	-18.75	-EEE-STTEEEESTTEEEE-

Table 2.2: $^{(*)}$ [kcal/mol]. Same as table 2.1 for the coarse graining based on the strings of secondary structure SSS[8].

i	ΔE_i $^{(*)}$	$\sigma(E_i)$ $^{(*)}$	P_i	$T\Delta h_i$ $^{(*)}$	$T\Delta S_i^{\mathrm{b}}$ $^{(*)}$	ΔG_i $^{(*)}$	Mesostate
1	-3.9	9.13	0.3232	-3.6	2.65	-6.60	000021000000210000
2	-1.6	9.36	0.0286	-2.6	5.91	-7.54	000021000000210010
3	-1.7	9.18	0.0274	-1.9	3.34	-5.08	000111000000210000
4	-1.6	9.06	0.0226	-1.4	1.65	-3.28	000021000001110000
5	-0.9	9.40	0.0211	-2.9	6.50	-7.44	000011000000210000
6	0.1	9.58	0.0174	-2.6	9.09	-8.98	000021000000200000
7	-3.4	9.56	0.0124	-1.8	8.68	-12.03	010011000000210000
8	-1.9	9.46	0.0087	-1.5	7.36	-9.26	000020000000210010
9	-0.2	9.05	0.0077	-2.0	1.54	-1.76	100021000000210000
10	1.5	9.81	0.0072	-1.5	12.47	-10.99	000020010000210000
11	0.5	9.28	0.0070	-2.1	4.77	-4.28	000021000000210001
12	-0.6	9.36	0.0065	-2.4	5.80	-6.42	000001000000210000
13	-1.2	9.63	0.0064	-1.2	9.79	-11.03	000021000000200001
14	-0.1	9.24	0.0057	-2.4	4.24	-4.35	001021000000210000
15	1.7	10.22	0.0053	-2.6	18.60	-16.87	000020000000210000
16	-1.3	9.42	0.0052	-2.6	6.77	-8.10	000121000000210000
17	-2.3	10.38	0.0051	-2.5	21.17	-23.52	010021000000210000
18	-0.4	9.53	0.0045	-2.6	8.28	-8.63	000021000000110000
19	3.3	9.18	0.0044	-1.5	3.29	-0.00	000021000010200000
20	-1.5	9.29	0.0039	-2.4	4.84	-6.36	000021000001210000
21	-1.9	9.19	0.0032	-0.8	3.43	-5.38	000010000000210010
22	-7.9	8.69	0.0032	0.7	-3.23	-4.67	001121001000210010
23	-0.6	9.61	0.0031	-0.9	9.42	-9.97	000111000000210010
24	-7.7	9.13	0.0027	1.5	2.62	-10.31	001111001000210010
25	-0.9	9.41	0.0023	-1.1	6.63	-7.53	000021000000210011
26	1.7	9.19	0.0022	-1.9	3.48	-1.77	000011000000210010
27	-1.2	10.34	0.0021	-1.3	20.51	-21.75	010001000000210000
28	2.2	9.68	0.0020	-0.9	10.48	-8.31	000111000000200000
29	-3.6	8.93	0.0019	1.9	-0.13	-3.44	001111001000211000
30	0.6	8.89	0.0018	0.3	-0.60	1.18	000111000001110000
31	-0.2	9.33	0.0018	0.4	5.41	-5.65	000002000000210000
32	0.8	8.83	0.0017	-0.7	-1.49	2.27	001111000000210000
33	4.5	9.56	0.0016	-1.6	8.77	-4.26	000021000000100000
34	0.5	9.35	0.0016	-1.9	5.75	-5.21	000021000000010000
35	2.7	9.47	0.0015	-0.5	7.48	-4.77	000021000001100000
36	1.6	9.82	0.0014	-0.8	12.55	-10.96	000021000010010000
37	1.7	10.05	0.0014	-1.7	15.96	-14.27	001011000000210000
38	-3.6	8.75	0.0013	1.2	-2.53	-1.04	001121001000211000
39	1.4	9.07	0.0012	-1.1	1.88	-0.49	000020000000211000
40	1.6	9.06	0.0012	11.8	1.64	-0.04	010001111111111111
41	0.2	9.70	0.0012	-1.4	10.74	-10.53	000101000000210000
42	6.1	9.40	0.0012	-0.8	6.39	-0.33	010010000000210000
43	0.1	9.89	0.0011	0.8	13.64	-13.53	000100100000210000
44	3.7	10.31	0.0011	-1.9	20.09	-16.42	000011000000200000
45	3.2	9.98	0.0011	-0.8	14.88	-11.66	010011000000210010
46	2.5	9.18	0.0011	-0.4	3.31	-0.80	000021000001110010
47	3.4	9.77	0.0010	-0.3	11.76	-8.40	001020010000210000
48	2.4	9.58	0.0010	-1.1	8.96	-6.57	100021000000200000
49	1.5	9.45	0.0010	-0.7	7.11	-5.58	000011000001110000
50	4.8	9.47	0.0010	-0.5	7.37	-2.59	000021000010100000

Table 2.3: $^{(*)}$ [kcal/mol]. Same as table 2.1 for the coarse graining based on the strings of rotational states SRA[4]. The quantity Δh_i is the configurational entropy loss of a string considering all the contributes due to the string sites, as explained in section 2.5.1.

i	ΔE_i $^{(*)}$	$\sigma(E_i)$ $^{(*)}$	P_i	$T\Delta H_i$ $^{(*)}$	$T\Delta S_i^{\mathrm{b}}$ $^{(*)}$	ΔG_i $^{(*)}$
1	-4.5	8.94	0.2384	-6.1	8.95	-13.48
2	-3.5	9.00	0.0222	-4.5	9.72	-13.18
3	-2.8	9.00	0.0168	-4.4	9.68	-12.46
4	-2.7	9.06	0.0151	-4.3	10.47	-13.20
5	-1.6	9.10	0.0148	-4.3	11.04	-12.69
6	-0.9	9.14	0.0142	-4.3	11.61	-12.52
7	-2.6	8.93	0.0129	-4.2	8.74	-11.38
8	-2.4	9.00	0.0121	-4.1	9.65	-12.04
9	-3.9	9.09	0.0099	-4.0	10.95	-14.84
10	-2.8	9.01	0.0095	-4.0	9.84	-12.63
11	-2.5	8.97	0.0086	-3.9	9.35	-11.90
12	-4.7	8.83	0.0076	-3.8	7.35	-12.09
13	-7.4	8.99	0.0066	-3.7	9.61	-17.05
14	-4.1	8.89	0.0063	-3.7	8.23	-12.36
15	-1.4	9.20	0.0058	-3.7	12.52	-13.89
16	-1.8	8.98	0.0051	-3.6	9.41	-11.22
17	-5.7	8.88	0.0049	-3.6	8.13	-13.83
18	1.1	9.18	0.0047	-3.5	12.19	-11.12
19	-3.1	9.29	0.0046	-3.5	13.77	-16.92
20	-2.3	8.83	0.0042	-3.4	7.44	-9.73
21	-2.5	9.18	0.0037	-3.4	12.25	-14.78
22	-3.4	9.13	0.0036	-3.3	11.54	-14.98
23	-2.5	8.76	0.0035	-3.3	6.54	-9.04
24	-2.6	9.18	0.0031	-3.2	12.21	-14.83
25	-3.5	8.91	0.0031	-3.2	8.43	-11.95
26	-0.5	8.89	0.0030	-3.2	8.21	-8.71
27	-0.2	9.04	0.0029	-3.2	10.22	-10.42
28	-0.7	8.86	0.0026	-3.1	7.87	-8.53
29	-3.2	9.13	0.0026	-3.1	11.49	-14.69
30	-0.4	8.91	0.0025	-3.1	8.45	-8.86
31	-7.1	9.34	0.0024	-3.1	14.40	-21.50
32	-2.5	8.74	0.0023	-3.1	6.20	-8.73
33	-1.3	9.07	0.0023	-3.1	10.64	-11.94
34	-0.1	9.15	0.0020	-3.0	11.71	-11.81
35	-2.5	9.25	0.0020	-3.0	13.15	-15.66
36	-1.3	8.99	0.0019	-2.9	9.55	-10.87
37	-2.8	9.51	0.0019	-2.9	16.85	-19.61
38	0.4	9.04	0.0018	-2.9	10.25	-9.82
39	-1.2	8.90	0.0017	-2.9	8.40	-9.60
40	-1.2	8.78	0.0017	-2.8	6.79	-8.00
41	-3.8	8.81	0.0016	-2.8	7.12	-10.95
42	-1.8	9.40	0.0016	-2.8	15.32	-17.09
43	-3.5	8.35	0.0016	-2.8	1.18	-4.65
44	-0.5	8.36	0.0014	-2.7	1.28	-1.83
45	1.2	9.26	0.0014	-2.7	13.25	-12.00
46	-3.7	8.59	0.0013	-2.7	4.22	-7.87
47	-2.0	9.12	0.0013	-2.7	11.40	-13.35
48	-0.7	9.28	0.0012	-2.6	13.62	-14.29
49	-2.2	9.46	0.0012	-2.6	16.16	-18.38
50	-2.1	9.11	0.0012	-2.6	11.21	-13.26

Table 2.4: $^{(*)}$ [kcal/mol]. Same as table 2.1 for the clustering based on Cα-RMSD with cutoff 1.5 Å.

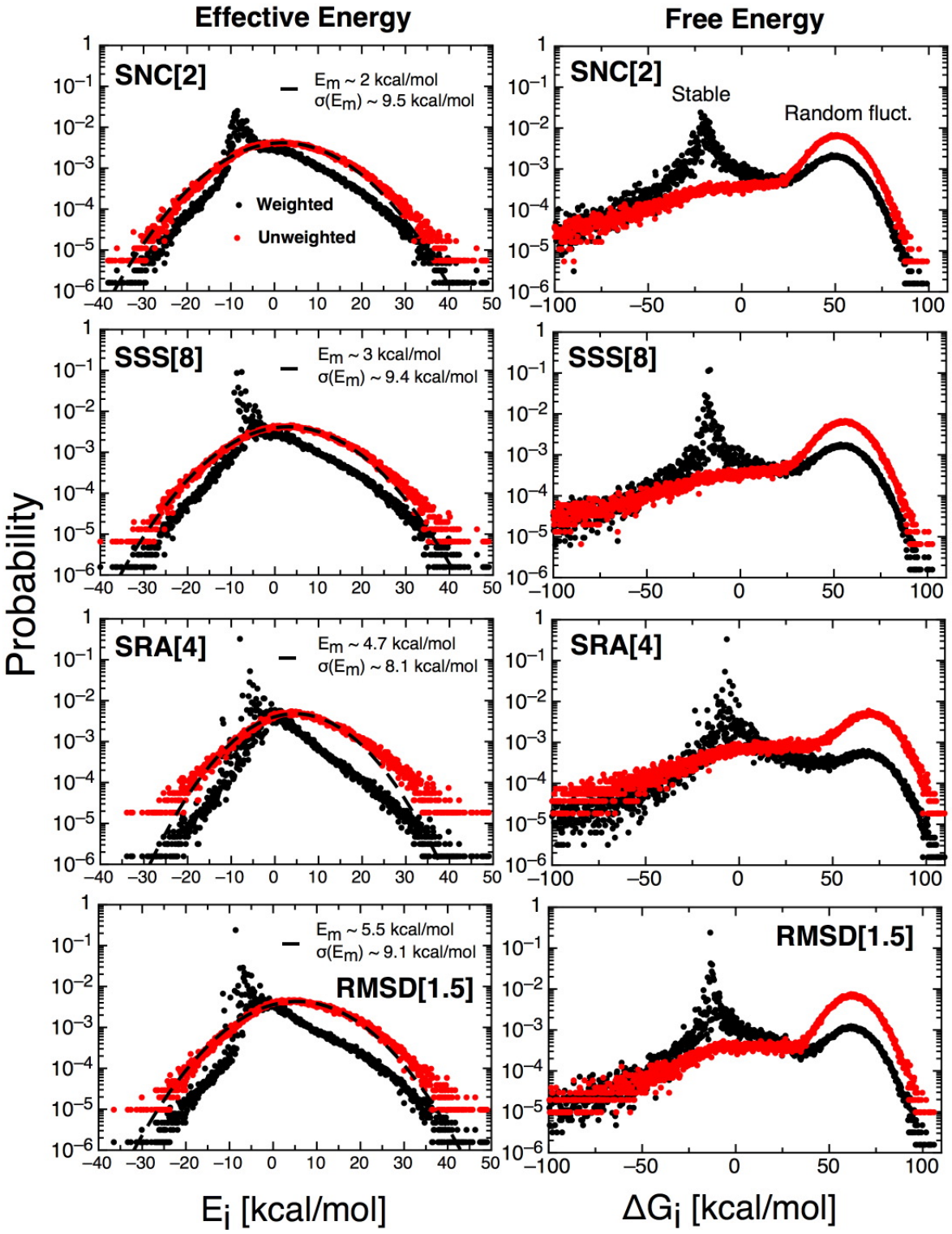

Figure 2.4: The energy (left column) and the free energy (right column) distributions among the ensemble of mesostates: black points are weighted with the mesostate probabilities while the red ones are not. Peaks in the weighted energy distributions corresponds to the folded mesostate. The unweighted energy distributions follows always a gaussian. The weighted free energy distributions are double peaked corresponding respectively to the folded and the unfolded state.

The probability of a mesostate is then

$$P_i = \frac{1}{M}\sum_{t=1}^{M}\theta_i(t) = \frac{n_i}{M} \tag{2.31}$$

where n_i is the observed frequency of the mesostate ω_i and M is the total number of microstates in the time series. In tables 2.1, 2.2, 2.3 and 2.4 the first 50 most populated mesostates with their thermodynamic parameters ($\Delta E_i = \overline{E_i} - \overline{E}$, $T\Delta H_i = T(H_i - H)$, $T\Delta S_i^b = T(S_i^{\mathrm{b}} - S^{\mathrm{b}})$ (where $S^{\mathrm{b}} = \sum_i P_i S_i^{\mathrm{b}}$), $\Delta G_i = \Delta E_i - T\Delta S_i^{\mathrm{b}}$) are shown. The mesostates having $\Delta G_i < 0$ represent the stable mesostates and those with $T\Delta S_i^{\mathrm{b}} < 0$ are mesostates not favored by the internal entropy but only by the total (or effective) energy, an example are single event mesostates, namely mesostates encountered just once along the trajectory. Thus the ΔG_i is a quantity that gives accounts on the statistical stability of the mesostates, whether or not the mesostates are statistically significant or merely fluctuations.

In figure 2.4 the weighted and unweighted distributions for the energies $\overline{E_i}$ and free energies $\Delta G_i = \Delta E_i - T\Delta S_i^{\mathrm{b}}$ (computed according to the relation 2.16) among the ensemble of mesostates are shown. The energy distributions give accounts on how the enthalpy is distributed among the mesostates. The unweighted energy distributions follow a gaussian in all the cases with mean E_m ranging from 2 kcal/mol for the SNC[2] to 5.5 kcal/mol for RMSD[1.5]; for the weighted distributions a pronounced peak at ~ -9 kcal/mol it is present. The peak correspond to the effective energy of the folded mesostate. In all the cases the right tails of the weighted energy distributions are gaussian because high energy mesostates are randomly distributed , while the left tails are exponential because low energy mesostates follow the Boltzmann statistics. The energy peak $E_{\mathrm{fold}} \sim -8 - 9$ kcal/mol is about twice the mean energy $\overline{E} = -4.04$ kcal/mol for all the mesoscopic descriptions as it can be seen from the tables 2.1, 2.2, 2.3 and 2.4. The mesostates having $\overline{E_i} < \overline{E}$ are enthalpy driven (their stability is governed by the energy) while those having $\overline{E_i} > \overline{E}$ are entropy driven (their stability is governed by the fluctuations). The free energy distributions (second column in figure 2.4) give accounts of the interplay between the internal entropies (vibrational modes) and the energies among the ensemble of mesostates. In particular the weighted distributions show a double peak which corresponds to the ensemble of stable and unstable mesostates respectively. Within the former the folded mesostates can be found. The stable peak is quite narrow in the folded state and the free energy contributions are both due to low enthalpy and high entropy due to the large energy fluctuations. The unfolded peak is gaussian and it is characterized by mesostates having low energy fluctuations (close to 0), but because of their large degeneracy in number their overall contribution to the distribution is essentially due to the conformational entropy that scales as $k_B \ln N$. A consequence of this interpretation of the second peak is that all the mesostates contributing to that are essentially indistinguishable and behave as configurational microstates. Finally, the two peaks define two *macrostates* in a thermodynamical sense and their boundaries in terms of the free energy G_i weakly depend on the kind of mesoscopic description adopted. The unweighted free energy distributions maintain the unfolded gaussian peak while the folded peak disappears to be replaced by long left tail. The unweighted distributions merely tell that there are many mesostates with high free energy and very few having low free energy.

2.4.2 Disorder and complexity of the ensemble of mesostates

Can we say something about the quality of the mesoscopic descriptions adopted? Let's consider the information H. Prior to the coarse graining procedure, the number of M microstates corresponds to the total length of the simulation. After the coarse graining we obtain N mesostates with their probability distribution. It is worth to compute the information gain of the procedure

$$\Delta H_{\text{gain}} = H_{\max} - H \tag{2.32}$$

where $H_{\max}$ is the information of M initial microstates

$$H_{\max} = k_B \ln M \tag{2.33}$$

and H is the actual information of the N mesostates (Eq. 2.8). Shannon entropy H measures the information of an ensemble of states and since it is upper bounded, the difference 2.32 evaluates the information extracted from the microstates due to the coarse graining or clustering. In a more physical language, this quantity measures the *distance* of the ensemble of mesostates from their micro-canonical equilibrium which is reached when all the mesostates are equally probable [Ebeling and Klimontovič, 1984]; the smaller is this difference the more equally accessible are the mesostates. Conversely large values means that few mesostates are very populated out of the whole available ensemble. Yet, the difference ΔH_{gain} depends on the number of mesostates N sampled in the simulation. That because entropy is an extensive quantity that grows with the size of the system too. To get rid of the dependence on the system size, we define the statistical disorder of the ensemble of mesostates as

$$\begin{aligned} \mathcal{D}(H) &= \frac{H}{H_{\max}} \\ &= \frac{H}{k_B \ln N} \end{aligned} \tag{2.34}$$

This quantity was first introduced by Landsberg [Landsberg, 1984] to decouple the disorder from the entropy. Disorder $\mathcal{D}$ is an intensive quantity such that $0 \leqslant \mathcal{D} \leqslant 1$. If $\mathcal{D} = 0$ only one mesostate is populated while for $\mathcal{D} = 1$ all the mesostates are equally populated meaning that the disorder is maximun. This quantity allows to compare different systems and different descriptions as it does not depend on the number of accessible mesostates, which is an advantage for systems having a growing number of states such as in a simulation. We define the same quantity for the entropies based on the energy fluctuations (mesoscopic), such that

$$\mathcal{D}(S) = \frac{S^{\text{m}}}{S^{\text{b}}} \tag{2.35}$$

where we have chosen as max entropy the bound entropy S^{b} for the simple reason that for a random choice of N mesostates, with $N \ll M$, S^{m} does not exceed S^{b}. Following Landsberg's arguments, from the definition of disorder one can define *order* by taking $1 - \mathcal{D}$. In particular in [Shiner et al., 1999] a definition of complexity has been suggested in terms of a product disorder-order such as

$$\mathcal{C} = \mathcal{D}(1 - \mathcal{D}) \tag{2.36}$$

The previous definition can be seen as a degree of structure (in a generalized sense) or self-organization exhibited by the system. Here we use it to compare different kind of description relatively to a certain

Description	N	$\Delta H_{\rm gain}/k_B$	H/k_B	$\mathcal{D}(H)$	$\mathcal{C}(H)$	$S^{\rm m\ (*)}$	$S^{\rm b\ (*)}$	$\mathcal{D}(S)$	$\mathcal{C}(S)$
SNC[2]	181214	3.5	9.5	0.80	0.159	0.1325	0.1350	0.98	0.017
SSS[8]	152237	5	8	0.67	0.219	0.1213	0.1458	0.83	0.139
SRA[4]	54364	7	6	0.54	**0.248**	0.0838	0.1833	0.46	**0.248**
RMSD[1.5]	102515	6	7	0.61	0.237	0.1106	0.1565	0.70	0.207
RMSD[2.0]	43547	7.5	5.5	0.54	**0.248**	0.0883	0.1788	0.49	**0.249**
RMSD[2.5]	15552	8.5	4.5	0.48	**0.249**	0.0731	0.1941	0.37	0.234
RMSD[3.0]	4921	9.5	3.5	0.44	0.246	0.0621	0.2050	0.30	0.211
RMSD[3.5]	1187	10	3	0.43	0.245	0.0543	0.2128	0.25	0.190
RMSD[4.0]	401	10.5	2.5	0.41	0.241	0.0450	0.2221	0.20	0.161

Table 2.5: $^{(*)}$ [kcal/mol/K]. Overview of the entropies for all the descriptive methods adopted to coarse grain the conformation space of the GSGS. $\Delta H_{\rm gain}$ is the information gain due to the coarse grain and $\mathcal{D}$ is its statistical disorder and $\mathcal{C}$ the complexity.

system, so that it can be seen as an optimization measure on the kind of description. From the definition (Eq. 2.36), $\mathcal{C}$ has a maximum for $\mathcal{D} = 0.5$ that is $\mathcal{C} = 0.25$. It is interesting to compare the complexity values for the two definitions of disorder provided, one purely informational 2.34, the other mesoscopic 2.35. In table 2.5 the computed values for $\Delta H_{\rm gain}$, H, $S^{\rm m}$, $S^{\rm b}$ as well as the informational and mesoscopic disorder and complexity are reported. The sum of $S^{\rm m}$ and $S^{\rm b}$ gives $S = 0.2648$ kcal/mol/K. Combining the informational and mesoscopic complexities as an optimization measure of the descriptions, the optimal complexities are obtained for the descriptions based of rotational strings and RMSD[2.0] (bold character in table). Complexities can be viewed as a sort of efficiency measure for the coarse graining, the higher it results, the more optimal is the redistribution of the total entropy among the vibrational $S^{\rm b}$ and mesoscopic $S^{\rm m}$ parts, in particular when $S^{\rm m}/S^{\rm b} = H/H_{\rm max} = 0.5$.

2.4.3 Non-convergence of the mesostates

The total number of visited mesostates N is a function of the sampling size M and its dependency over the time is equivalent to a vocabulary that grows in a text, the equivalence being made between mesostates and words. Scrolling a text the number of known words increases the vocabulary size and grows sub-linearly with the total number of scrolled words (see [Kornai, 2002] and references therein). The sub-linearity, that means $N \propto M^{\alpha}$ with $\alpha < 1$, reflects the non-bernoullian (i.e. non-random) nature of a real text. For the sampling of mesostates the process is similar. In figure 2.5 (A) we show how the number of explored mesostates grows with the sampling size depending on the different coarse graining procedures adopted. In all the examples the mesostates grow sub-linearly with a different exponent while the rate to visit a new mesostate N/M (B) decreases slowly. Note that N/M is $dN/dM \propto \alpha M^{\alpha-1}$. For α close to 1 the decay of the rate is very slow: the slowest is that for the RMSD[1.5] ($\propto 1/M^{0.1}$). Much faster are the decays for RMSD[2.5] and RMSD[3.0], because the large cutoffs recovers exhaustively the whole conformational spaces sampled by the peptide. Thus, this means that the total number of possible mesostates $N^* \lesssim M$. The rest of the descriptions show a slow decay, that is $\propto 1/M^{0.2}$. The slow decay then implies $N^* \gg M$, where N^* is the max number of possible mesostates.

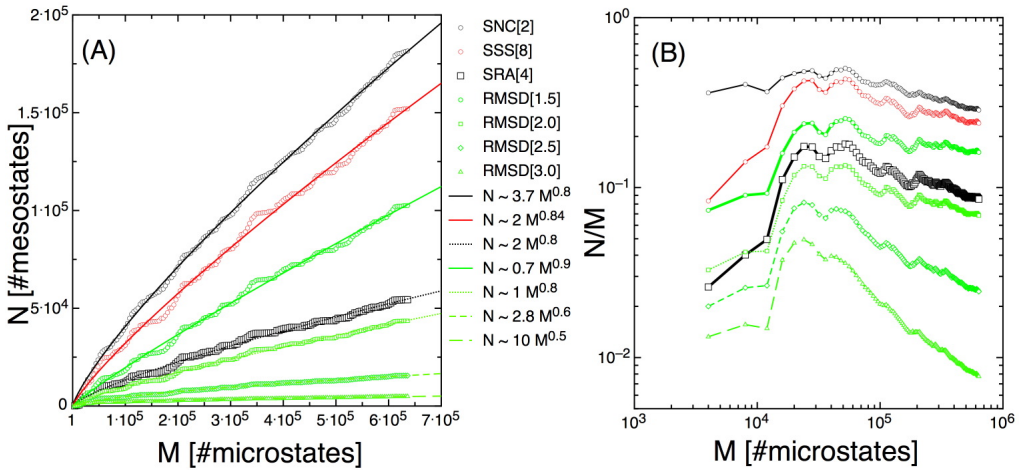

Figure 2.5: (A) The sublinear increasing of the number of visited mesostates N as a function of the number of sampled microstates M for different coarse graining procedures; (B) the ratio N/M which is the probability to sample a new mesostate over the time.

To quantitatively understand the sub-linearity it is worth to notice that visiting new mesostates is in principle equivalent to an ideal Bernoulli process in which the random variable is $\vartheta = 0$ when a new visited microstate is included into a previously visited mesostate and $\vartheta = 1$ otherwise. Let's call p_1 the probability to have $\vartheta = 1$. The counter function $\vartheta(t)$ is such that one can write

$$N(M) = \sum_{t=1}^{M} \vartheta(t) \tag{2.37}$$

which is exactly what is shown in figure 2.5 (A). Let's now assume the process bernoullian, thus given M microstates the probability to observe n mesostates is merely given by the binomial distribution

$$\mathcal{P}(n, M) = \binom{M}{n} p_1^n (1 - p_1)^{M-n} \tag{2.38}$$

which has the mean value

$$N(M) = p_1 M \tag{2.39}$$

Thus for a bernoullian process N, the sampled number of mesostates, grows linearly with the sampling size M, where the total number of possible mesostates is always $N^* \gg M$ ([2]). We argue that the process of collecting new mesostates is strictly non bernoullian since the probability p_1 decreases with the sampling size as shown in figure 2.5 (B). Therefore a feature of the behaviour of $N(M)$ is then the lack of convergence in the number of mesostates, in particular the sub-linear increasing demonstrates that there is no hope to obtain from a simulation an exhaustive sampling of the space of mesostates, as already noticed in [Cavalli et al., 2003]. However there should be something intrinsic connecting of conformational free energy landscape features and the power law decreasing of the emission rate N/M.

[2] $N^* = \lim_{M\to\infty} N(M)$, and the sampling size M can never be of the order of N^*

2.4.4 Rank ordered distributions and density of mesostates

The index i in equation 2.31 orders the mesostates by decreasing probabilities so that P_1 is the probability of the most populated mesostate, P_2 the second most populated mesostate and so on. Thus P_i is a rank-ordered distribution of probabilities as well as the set of f_i frequencies and i gives the rank r. On the other hand the rank r is a function of the probability P which gives the total number of mesostates that have probability greater equal than P. This number can be written as

$$r(P) = \sum_{P' \leq P} k(P') \tag{2.40}$$

where the function $k(P)$ gives the number of distinct mesostates that share the same probability P. The function $k(P)$ represents then the density of mesostates given the normalization

$$\sum_{P} k(P) = N \tag{2.41}$$

and

$$\sum_{P} Pk(P) = 1 \tag{2.42}$$

In figure 2.6 we show the calculated densities $k(P)$ for all the descriptions used to create mesostates out of the simulation of the GSGS. The densities of mesostate calculated from the data are shown in Figure 2.6 for all the descriptions employed. All the densities show a cross over area that corresponds approximatively to a probability range $P_c \sim 10^{-4} - 10^{-3}$; after the cross over a plateau is reached which means that with high probability there is only a single mesostate. The folded state is obviously the last point of the curves. The decreasing parts until the cross over follow a power law decay $k(P) \propto (P_c/P)^D + 1$ with $D \sim 2$ for the strings (A) and a log-normal for the RMSD clustering and the SRA[4],

$$k(P) \sim \frac{N}{\sqrt{2\pi}\sigma} \frac{1}{P} e^{-(\ln P - \mu)^2/2\sigma^2} \tag{2.43}$$

though an overall decay $k(P) \propto 1/P^D$ with $D \sim 2$ can be recognized. The high number of unassigned microstates is a signature that the used clustering algorithm is at the first neighbors. Clusters are built up around certain microstates that are centers as they have the highest number of first neighbors in RMSD with respect to a fixed cutoff. The cutoff defines a size in the conformation space, which introduces finite size effects in the clustering generating artifacts, the unassigned microstates. To see that let us consider a two dimensional plane as represented in figure 2.7 and let the black spots be microstates. We want to collect them according to a distance criteria. Given the distance we first introduce a small cutoff and then we group them into the clusters represented by the white filled circles whose radius is the chosen cutoff. It is clear the clusters cannot recover the whole space in which the microstates are distributed and some microstates remain as spurious unassigned microstates. Considering a larger cutoff the spurious microstates do not disappear completely but decrease in absolute value as the radius of the clusters increases. Thus the spurious microstates are just a finite size effect, due to the fact that clustering introduce a finite size in a continuous conformational space. A multi-neighboring algorithm would eliminate them by including into larger clusters defined at second neighbor, third neighbor and so on until convergence. Such an approach, although quite accurate, is computationally very expensive (see for example [Johnson, 1967]).

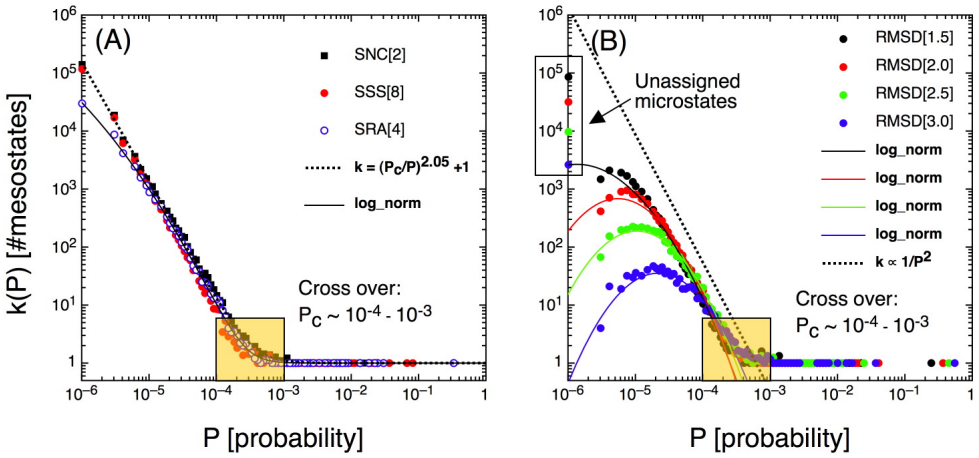

Figure 2.6: The density of mesostates of the GSGS from the descriptions based on strings (A) and RMSD clustering (B).

The log-normal distribution It is not surprising the fact that we observe a log-normal distribution. One obtains a log-normal distribution of a random variable when the logarithm of the variable is gaussian distributed. We have already noticed that the energy density for the ensemble of microstates (see equation 2.44) follows a gaussian and as well as for the ensemble of mesostates (figure 2.4). We want to express the equation 2.44 in terms of probabilities $P = 1/Ze^{-\beta E}$. Let's consider the energy density

$$k(E) = \frac{1}{\sqrt{2\pi}\sigma_E} e^{-(E-\overline{E})/2\sigma_E^2} \tag{2.44}$$

thus from the cumulative $\int k(P)dP$ the density turns out to be

$$k(P) = \frac{1}{\sqrt{2\pi}\sigma_E} \frac{1}{P} e^{-(\ln P + \ln Z + \beta\overline{E})^2/2\beta^2\sigma_E^2} \tag{2.45}$$

By taking the $\ln$ and using the relation $S/k_B = \ln Z + \beta\overline{E}$ we obtain

$$\begin{aligned} \ln k(P) &= -\ln P - \frac{(\ln P + \ln Z + \beta\overline{E})^2}{2\beta^2\sigma_E^2} + \text{constant} \\ &= -\frac{(\ln P)^2}{2\beta^2\sigma_E^2} - \left(\frac{S}{k_B\beta^2\sigma_E^2} + 1\right) \ln P + \text{constant} \end{aligned} \tag{2.46}$$

which is quadratic in $\ln P$ with a maximum in

$$P^* = \frac{1}{Z} e^{-\beta\overline{E} - \beta^2\sigma_E^2} = e^{-S/k_B - \beta^2\sigma_E^2} \tag{2.47}$$

That means that the peak corresponds to the most probable microstate. The linear part has a slope

$$D \sim \frac{S}{k_B\beta^2\sigma_E^2} + 1 \tag{2.48}$$

The equations from 2.45 to 2.48 are referred to an ensemble of microstates having free energy G, mean energy $\overline{E}$ and entropy S. In the case of an ensemble of mesostates such as those of figure 2.6 we assume a gaussian free energy distribution

$$k(\Delta G_i) \sim \frac{N}{\sqrt{2\pi}\sigma_G} e^{-(\Delta G_i)^2/2\beta^2\sigma_G^2}$$

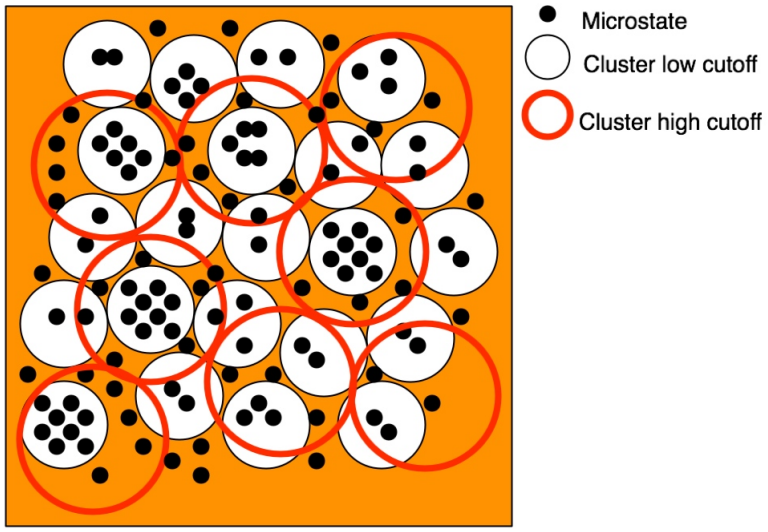

Figure 2.7: An example to explain how clustering based on first neighboring produces spurious unassigned microstates.

which, even though badly corresponds to the unweighted free energy distributions of figure 2.4, it is nevertheless useful for qualitatively purposes. For each mesostate we take $P_i \sim e^{-H^*/k_B} e^{-\beta \Delta G_i}$ so that we obtain in $\ln$ scale

$$\ln k(P_i) = -\frac{(\ln P_i)^2}{2\beta^2\sigma_G^2} - \left(\frac{H^*}{k_B\beta^2\sigma_G^2} + 1\right) \ln P_i + \text{constant} \tag{2.49}$$

For large enough $\beta\sigma_G$ is linear in $\ln P_i$ with a slope

$$D \sim \frac{H^*}{k_B\beta^2\sigma_G^2} + 1 \tag{2.50}$$

The distribution $k(P_i)$ has a maximum for $P^* = e^{-H^*/k_B - \beta^2\sigma^{2G}}$. Looking to the figure 2.6 only the points below the cross over P_c contribute to the distributions, namely the poorly populated mesostates (those that have $\Delta G_i > 0$). For this subset of mesostates we expect that the Shannon entropy shall be $H^* \sim k_B \ln N^*$ where N^* is a the theoretical number of accessible mesostates which is much larger than the effective number of mesostates $N(M)$ observed in the simulations. Thus given the method of coarse graining the density of mesostates allows us to estimate H^* and then the max number of accessible mesostates and moreover the fluctuations $\beta\sigma_G$ of the free energies G_i. We have fitted all the densities with the log-normal distribution and obtained the values reported in table 2.6 for the parameters H^*/k_B (N^*), $\beta\sigma_G$. We have also estimated the exponent D and the cross over P_c from the power-law fits

$$k(P) = (P_c/P)^D + 1 \tag{2.51}$$

for the mesoscopic descriptions SNC[2], SSS[8], SRA[4], which is always about 2.1. It is interesting to compare the estimated values of N^* obtained from the log-normal fits with those theoretical. For instance for the SNC[2] one has $2^{40} \sim 10^{12}$ possible string which roughly corresponds to the fittet value; for SSS[8] one has $8^{18} \sim 10^{16}$ which is 7 orders of magnitudes larger than the fitted value; for SRA[4] one

Description	H^*/k_B	N^*	$\beta\sigma_G$	D (power law fit)	P_c	r_c	P_{stable}
SNC[4]	~ 29	$\sim 4 \cdot 10^{12}$	~ 4.4	2.05	$3.4 \cdot 10^{-4}$	260	0.33
SSS[8]	~ 22	$\sim 5 \cdot 10^{9}$	~ 3	2.09	$2.3 \cdot 10^{-4}$	149	0.52
SRA[4]	~ 14.8	$\sim 2.7 \cdot 10^{6}$	~ 2	2.15	$2.1 \cdot 10^{-4}$	262	0.67
RMSD[1.5]	~ 11.6	$\sim 1.1 \cdot 10^{5}$	~ 1.1	lognormal	$2.5 \cdot 10^{-4}$	227	0.57
RMSD[2.0]	11.0	~ 60000	1	lognormal	$2.5 \cdot 10^{-4}$	255	0.65
RMSD[2.5]	10.4	~ 33000	1	lognormal	$2.5 \cdot 10^{-4}$	318	0.75
RMSD[3.0]	9.8	~ 18000	1	lognormal	$2.5 \cdot 10^{-4}$	368	0.87

Table 2.6: The parameters estimated from the log-normal fit (H^*/k_B, N^*, $\beta\sigma_G$) of the densities of mesostates and the exponent D estimated from a power law fit (eq. 2.51) mainly for the string based mesostates. The number of stable mesostates r_c and their cumulative statistical weight P_{stable} are also reported in table. The r_c number should be regarded as an order of magnitude rather as an exact quantity.

has $4^{18} \sim 6 \cdot 10^{10}$ strings that are 4 order of magnitude larger than the fitted value. The large differences between the theoretical and effective number of accessible mesostates for the secondary structure and torsional strings reflects the fact that many strings of this type are not physically possible. Interestingly, the free energy fluctuations are comparable with the thermal fluctuations for mesostates based on the RMSD clustering. That is likely due to the more macroscopic character of these kind of mesostates. Moreover the comparison between the estimated values of D from power low fits with those deduced from the excellent log-normal fits suggest the correctness of the log-normal hypothesis. Thus maximun value for the density $k(P_i)$ corresponds to $P^* \sim 1/N^*$ that means the overwhelming majority of mesostates have an extremely low probability. A further thing that should be noted is that the cross over in figure 2.6 does distinguish between two kinds of mesostates, those that are statistically unstable $P < P_c$ and those highly populated (statistically stable), which coincide with negative values of ΔG_i. How many are the statistically stable mesostates? Going from the density of mesostates to the ranking distributions may give the answer. Let's consider a continuos version of the equation 2.40

$$r(P) = \int_P^1 k(P')dP' \tag{2.52}$$

and let us take

$$k(P) \sim \left(\frac{P_c}{P}\right)^D + 1 \tag{2.53}$$

with $D = \frac{H^*}{\beta\sigma_G} + 1$ and P_c the cross over probability. We obtain then for $P < P_c$

$$r(P) \sim \frac{P_c}{D-1}\left(\frac{P_c}{P}\right)^{D-1} \tag{2.54}$$

which inverted gives

$$P(r) \sim \frac{V}{(r_c + r)^\eta} \tag{2.55}$$

with V a constant and $\eta = 1/(D-1)$. If we have $D \gtrsim 2$ then $\eta \gtrsim 1$ and $P(r) \sim V/(r_c + r)$ is the Zipf-Mandelbrot law that fits with the ranked data of figure 2.8. The number r_c is the number of sta-

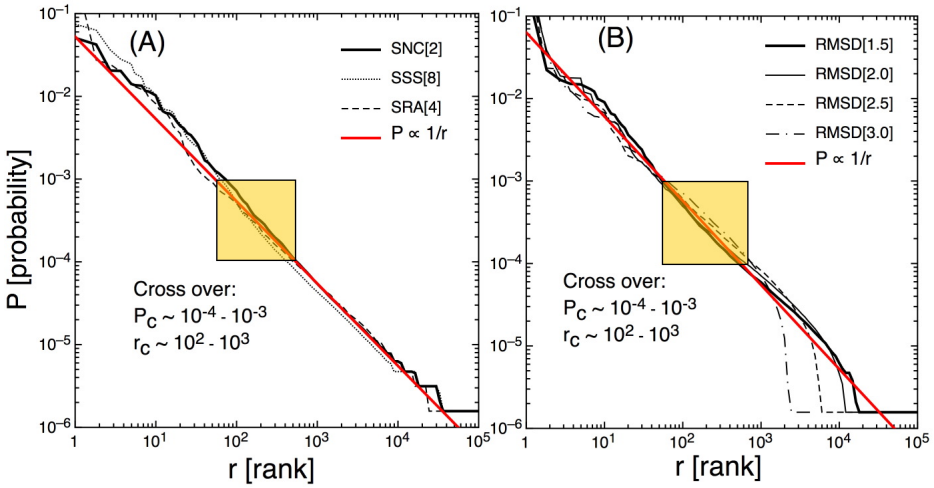

Figure 2.8: The ranking distributions for the GSGS from the descriptions based on strings (A) and RMSD clustering (B).

tistically stables mesostates. r_c values per description together with their cumulative statistical weight are reported in table 2.6. Zipf-Mandelbrot law simply states that if we have N species involved in a dynamical process the most used species have the double probability the be used than the second most used species, the triple probability to be used than the third most used species and so on. Again the example of a written text helps to understand. If one makes the statistics on a text by ranking the most used words usually the Zipf-Mandelbrot law is obtained with an exponent depending on the complexity of the corpus text. It has been observed (see for example [Kornai, 2002]) that highly complex texts usually exhibit an exponent $\eta \sim 1.2$. Zipf-Madelbrot law is usually interpreted as a signature of a hierarchy underling the process. In the case of a text the hierarchy is thought to be layered in about 3 levels: the first is given by the grammar rules of a language, the second is defined by semantics and the last provided by the style of the writer. In such a scheme the probabilities would decrease from the first to the last level since the grammar defines very common words, while a writer's style characterize special and possibly rare words. Going back to protein folding we have defined mesostates out of a myriad of microstates and their rank statistics have been done. A natural hierarchy places at the highest level the folded state and at the followings mesostates having higher energy and higher conformational entropy. At the lowest level there are mesostates whose density scales with $\sim 1/P^2$ and that are dominated by the entropy. Transition states or rarer mesostates, when viewed as ensemble, represent a massiv contribution to the conformation space. The main point of this treatment about the density of mesostates or the ranking distributions, is that only a very limited subset of mesostates robustly represent the conformation space and only those should be taken into account to describe it. That shall be formalized when the convergence of the sampled entropy will be addressed.

2.4.5 Structural interpretation of the mesostates

We have established that given a sampling of mesostates only a small subset is representative of the whole ensemble. In particular, from the density of mesostates and the rank ordered distributions (see

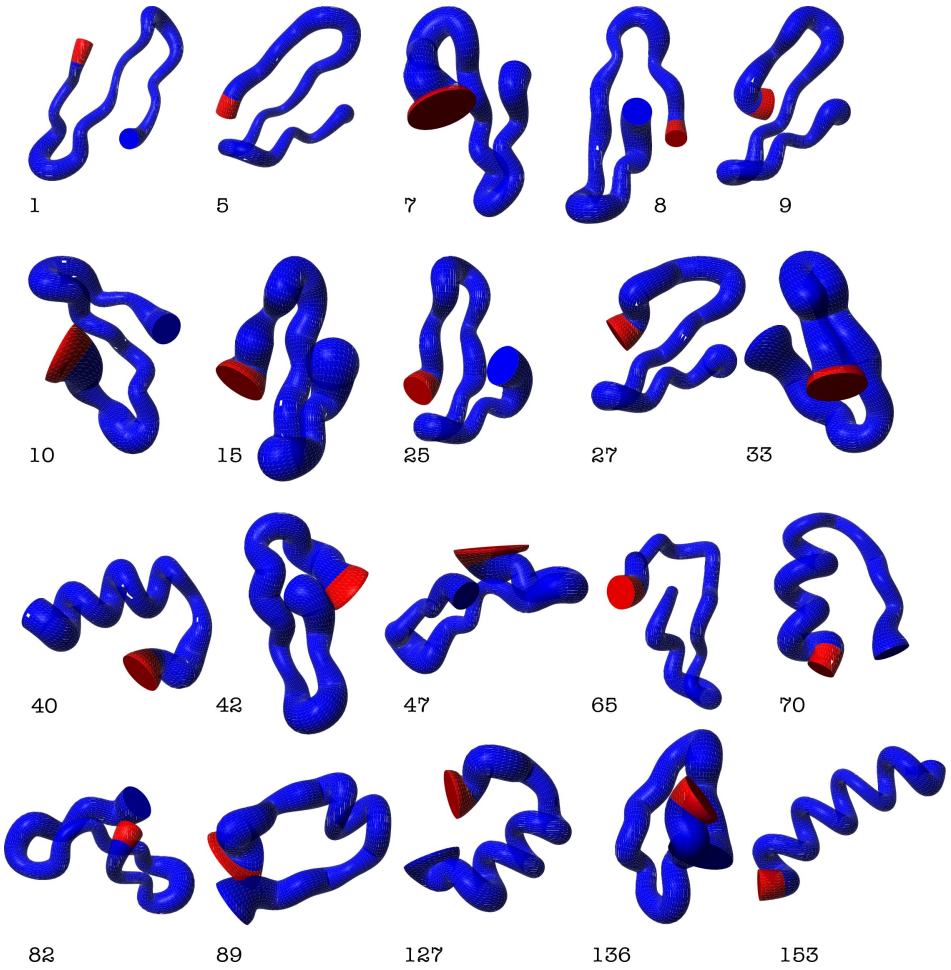

Figure 2.9: Selected mesostates corresponding to the description based on SRA[4] taken among the first $r_c = 262$ mesostates. The figures represent ensemble of structures with their fluctuations within a mesostate corresponding to a well defined string. The N-term of the polypeptide is colored in red. The structures are represented with their RMSF fluctuations by using the macro "sausage" implemented in program molmol.

figures 2.6 and 2.8) we found that this number corresponds to r_c (which gives the order of magnitude of the number of mesostates statistically well sampled), while the overwhelming majority of the mesostates follow a log-normal distribution. According to efficiency considerations, based on the measures of complexity, we choose the SRA[4] as our standard method of description of the system. Within the $r_c \sim 262$ mesostates we have selected some and represented with their structural characteristics in figure 2.9. In table 2.7 the values of energy, internal entropy and free energy differences are reported. The conformational landscape of the representative mesostates appears to be quite heterogeneous. Most of the mesostates selected are stable in the sense that their free energy difference is negative and among them some can be stabilized either by mainly the energy (the folded mesostate 1 is an example) or by the entropy (the mesostate 15 for example) or both (the mesostate 82). Unstable mesostates show $\Delta E_i > 0$ and $\Delta S_i^{\mathrm{b}} < 0$ as for example the helical-like state 127. The mesostates having $\Delta G_i \sim 0$ are those for which

Id number	SRA string	ΔE_i [kcal/mol]	$T\Delta S_i^{\mathrm{b}}$ [kcal/mol]	ΔG_i [kcal/mol]	P_i
1	000021000000210000	-3.9	2.65	-6.60	0.323
5	000011000000210000	-0.9	6.50	-7.44	0.021
7	010011000000210000	-3.3	8.68	-12.03	0.012
8	000020000000210010	-1.9	7.36	-9.26	0.009
9	100021000000210000	-0.2	1.54	-1.76	0.008
10	000020010000210000	1.4	12.47	-10.99	0.007
15	000020000000210000	1.7	18.60	-16.87	0.005
25	000021000000210011	-0.9	6.63	-7.53	0.002
27	010001000000210000	-1.2	20.51	-21.75	0.002
33	000021000000100000	4.5	8.77	-4.26	0.002
40	010001111111111111	1.6	1.64	-0.04	0.001
42	010010000000210000	6.1	6.39	-0.33	0.001
47	001020010000210000	3.3	11.76	-8.40	0.001
65	000111001000210010	-0.2	2.90	-3.09	0.0008
70	111111111111300000	3.3	1.21	2.11	0.0007
82	100021000100110000	-1.4	15.38	-16.74	0.0006
89	010001111111300000	5.7	5.95	-0.28	0.0006
127	010001111111111100	3.9	-2.65	6.63	0.0004
136	000020010000110000	2.8	18.34	-15.55	0.0004
153	111111111111111111	1.4	6.04	-4.59	0.0003

Table 2.7: The thermodynamic parameters defining the selected mesostates of figure 2.9. $\Delta E_i = \overline{E_i} - \overline{E}$, $T\Delta S_i^b = T(S_i^{\mathrm{b}} - S^{\mathrm{b}})$ (see equation 2.9 for the definition of S^{b}), $\Delta G_i = \Delta E_i - T\Delta S_i^{\mathrm{b}}$. Positive values of ΔG_i correspond to unstable mesostates (with positive ΔE_i and low $T\Delta S_i^b$). Among the unstable mesostates the bad sampled ones can be found. Positive values of ΔG_i are distributed according to a Gaussian (see figure 2.4). Mesostates with negative ΔG_i are stable and well sampled: they posses either low enthalpy or large vibrational entropy.

the energy loss is compensated by the same amount of internal entropy gain. Examples of them are the helical mesostates 40, 89 or the exotic mesostate 42 where the 2nd structure is native like while the the overall topology is not. We might speculate that among these mesostates the transition mesostates for folding are hidden, though this conclusion would causally be related to the description it originates from.

2.5 The organization of an ensemble of strings

2.5.1 Entropy and disorder of strings

The multidimensional characteristics of the mesostates defined by string based descriptors have not been used yet. We assumed each string site to variate within an alphabet of λ symbols so that a string $\mathbb{S}$ of length R ($R = 18$ for the methods SSS[8], SSS[4] and SRA[4] in the GSGS polipeptide) one has

$$\mathbb{S} = \mathbb{S}_1\mathbb{S}_2...\mathbb{S}_R \tag{2.56}$$

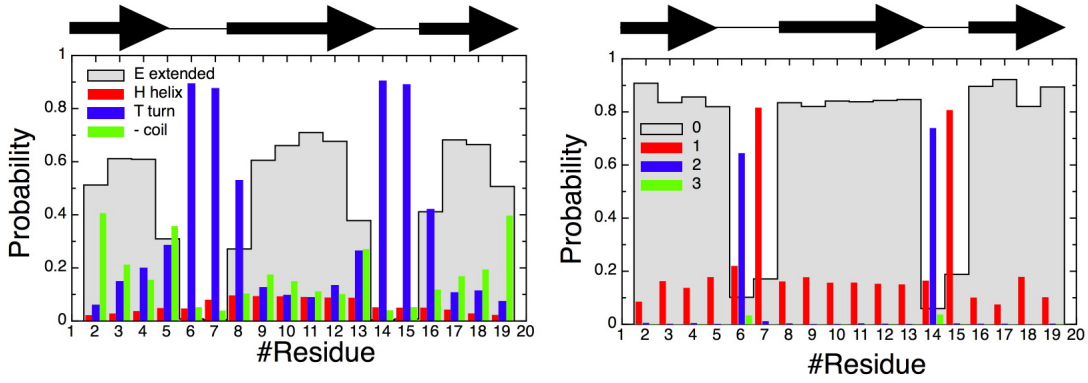

Figure 2.10: The string site probabilities for the mesostates based on SSS[4] (left) and SRA[4] (right) of the GSGS. The similarities between the two descriptions are evident.

where a site $\mathbb{S}_i$ assumes a symbol s in a set $(s_1, s_2, \cdots, s_\lambda)$. Thus from the simulations we can estimate the probability of a symbol in a string site through the identity

$$p(\mathbb{S}_i) = \frac{n(\mathbb{S}_i)}{M} \tag{2.57}$$

where M is the total number of microstates, $n(\mathbb{S}_i)$ is the frequency of a symbol in a string site i and with the normalization

$$\sum_{k=1}^{\lambda} p(\mathbb{S}_i = s_k) = 1 \tag{2.58}$$

for each string site i. For simplicity we write $p(\mathbb{S}_i = s) = p_i(s)$. The values of the string site probabilities per symbol for the mesostates based on the reduced secondary structure strings SSS[4] and the rotational strings SRA[4] are graphically shown in figure 2.10. The distributions reveal the strong beta propensity of the GSGS, although a not negligible amount of helix structure also characterizes the configurational landscape of the polipeptide. Given the string site probabilities we can define a combined probability for a given specific string $\mathbb{S}$ such as

$$p(\mathbb{S}) = \prod_{i=1}^{R} p_i(s) \tag{2.59}$$

from which the global normalization follows

$$\sum_{\mathbb{S}} p(\mathbb{S}) = \prod_{i=1}^{R} \left(\sum_{k=1}^{\lambda} p_i(s_k) \right) = 1 \tag{2.60}$$

The equation 2.59 is of course not fully correct since a certain amount of cross correlations characterize the chain of symbols in a string, as it represents a polipeptide chain. Nevertheless, since our aim is to provide a *description* of folding and not to construct a first principle model of it, we assume the probability 2.59 as an observable quantity of the ensemble of strings. After having computed the combined probabilities for the ensemble of strings SRA[4] observed in the simulation of the GSGS we have calculated their density distribution, which has been also weighted with the observed probabilities of the ensemble of mesostates. In figure 2.11 both the distributions are shown. The black distribution is the weighted one, it monotonically grows with the combined probability, namely the highest value of the

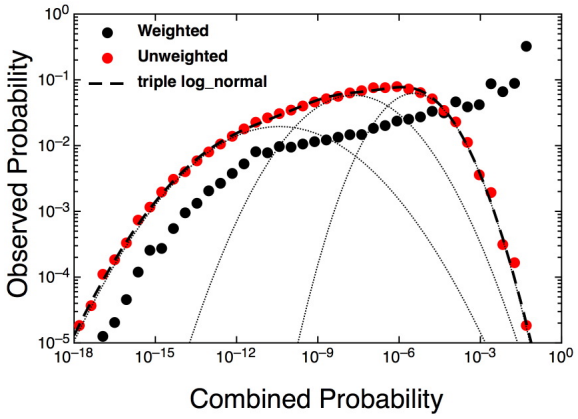

Figure 2.11: The distributions of the combined probabilities for SRA[4] in a log-log plot. The black dots are the observed string probabilities as a function of the combined probabilities while the red curve is the density of combined probabilities. The latter fits very well with triple log-normal distribution (dashed curve) which suggests that the ensemble of strings is organized in three sub-ensembles. The dotted curves are single log-normal distributions with the parameters estimated from the triple fit.

observed string probability corresponds to the highest value of the combined string probability (theoretical). The unweighted distribution (red data points) fits very well with a triple log-normal distribution (dashed curve in figure) which suggest that the ensemble of strings are organized in three sub-ensembles of strings. As we have seen previously the lognormal fit allow to estimate the maximal Shannon entropy of the ensemble of mesostates. From the maximal Shannon entropy, by simply taking its exponential, one obtains the mean maximal number of accessible mesostates. From the triple log-normal fit we obtain three values of entropies one for each sub-ensemble: $H_1/k_B \sim 12 \pm 3$, $H_2/k_B \sim 17 \pm 7$ and $H_3/k_B \sim 24 \pm 8$ respectively. The estimated errors from the fit are quite large whereas one can estimate the order of magnitude of the numbers of accessible strings for each macro-phase: we have $N_1 \sim 2 \cdot 10^5$, $N_2 \sim 10^7$ and $N_3 \sim 10^{10}$ respectively and the sum is dominated by N_3. This number is about one order of magnitude less than the theoretical upper limit $4^{18} \sim 7 \cdot 10^{10}$ for the string based description SRA[4]. This means that in theory all the strings of the description based on the discrete torsional angles are in principle accessible. Let us investigate the Shannon entropy associated to the probability 2.59. We first the entropy per string site (that is for SSS[8], SSS[4], SRA[4] the entropy of a residue) as

$$h_i = -k_B \sum_{k=1}^{\lambda} p_i(s_k) \ln p_i(s_k) \tag{2.61}$$

Since entropy is an extensive quantity, summing up along the string chain we obtain the total string entropy

$$h = \sum_{i=1}^{R} h_i \tag{2.62}$$

The two entropies h and H (defined in equation 2.8) assume the same value if and only if all the string sites are independent and in particular one has

$$H \leqslant h \tag{2.63}$$

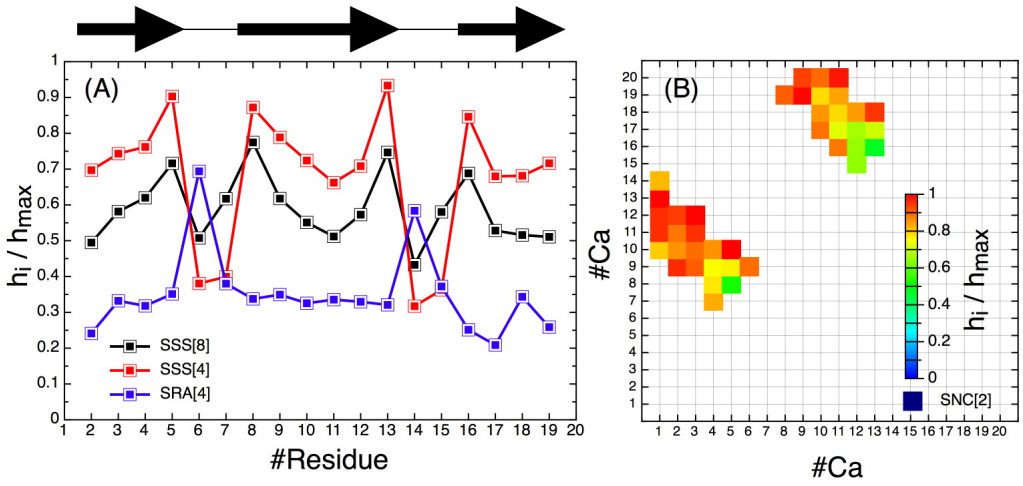

Figure 2.12: (A) Disorder per residue for the string based on secondary structure SSS[8], SSS[4] and on rotational angles SRA[4]; (B) the map of native contacts whose points represent the disorder of a Ca contact on which the strings SNC[2] are based.

which means that cross correlations have the overall effect to reduce entropy. Thus the entropy difference $h - H$ evaluates the mean information difference between a string considered as a *whole* and a string considered as a combination of independent string sites or in other words, $h - H$ measures the information due to the cross correlations between sites (that in the case of SSS and SRA represent the correlations between residues). However the difference $h - H$ strongly depends on the sampling size as the entropy H intrinsically grows with the sampling size M. Thus, it is natural to define a disorder $\mathcal{D}(h_i)$ per site and the string disorder $\mathcal{D}(h)$ as

$$\mathcal{D}(h_i) = \frac{h_i}{h_{max}} \tag{2.64}$$

and

$$\mathcal{D}(h) = \frac{1}{R}\sum_{i=1}^{R}\mathcal{D}(h_i) \tag{2.65}$$

where $h_{max} = \ln\lambda$ with λ the alphabet length. In figure 2.12 the plots of the disorder for each string based coarse graining are represented. In 2.12 (A) the string site disorders for the descriptions SSS[8], SSS[4] and SRA[4] are represented. The peaks present in the SRA[4] disorder profile correspond to the glycines 6 and 14 which notably correspond to a disorder minima in the profiles based on secondary structure. This is an apparent contradiction due to the fact that these glycines belong to the beta turns. In the secondary structure code a turn-like configuration precisely corresponds to a letter `T` (beta turn) for SSS[8] and `T+B` (beta turn and bend) for SSS[4]; while for strings based on rotational angles there can be many torsional local arrangements giving a glicine-serine loop stabilized by backbone hydrogen bond [Wilmot and Thornton, 1988]. Thus, it results evident that what can appear structured for a description might turn out to be unstructured in another and, yet, subjectivity and convenience of the observer become an important matter. In the (B) part of figure 2.12 a Cα contact map from which the strings of native contacts have been built is shown. The color shades in figure corresponds to the disorder linked to a specific Cα native contact. The contact map has the typical β-sheet structure coming from

long range (sequence distant) contacts. The disorder pattern reveals that the less disordered Cα contacts are those corresponding to the glycine-serine beta-turns: notably the contacts Asp5-Thr8 and Asp13-Thr16. All the long range contacts are very disordered (values close to 1) suggesting that they are rapidly created and disrupted in the dynamics. The SNC[2] disorder is then consistent with that from secondary structure SSS[8] and SSS[4] as both the descriptions strongly depend on the hydrogen bond network of the polipeptide. SRA[4] description is substantially different from the others since it does depend only on unrelated coarse grained degrees of freedom of the polipeptide. The values of string disorder are respectively 0.84 for SNC[2], 0.58 for SSS[8], 0.67 for SSS[4] and 0.35 for SRA[4].

The entropy h can be written as the usual average $h = \sum_{\mathbb{S}} p(\mathbb{S})h(\mathbb{S})$ with $h(\mathbb{S}) = -k_B \ln p(\mathbb{S})$ so that $h(\mathbb{S})$ is the entropy of a specific string. The values of $h(\mathbb{S})$ computed on the ensemble of observed strings rank these from the lowest value to the highest in a similar but more appropriate way as the entropy $H_i = -k_B \ln P_i$ where P_i is the estimated (from the simulations) probability of the i-th mesostate. If we assume that $h_{unf} = Rh_{max} = Rk_B \ln \lambda$ is the entropy of the unfolded state, the state such that all the strings are equally accessible, then given a string $\mathbb{S}$ the entropy difference $\Delta h(\mathbb{S}) = h_{unf} - h(\mathbb{S})$ has again a double interpretation: on one hand it represents the information gain of the string $\mathbb{S}$ with respect the unfolded state (namely the amount of information to encode a specific string); on the other hand $\Delta h(\mathbb{S})$ represents the conformational entropy loss from the unfolded state to a specific string. Clearly the folded string has the highest entropy loss with respect to the unfolded state.

2.5.2 String hierarchy and configurational hierarchies

The multidimensional character of the strings allows us to investigate weather a hierarchy of states exists or not in the configurational space. According to the conformational entropy $h(\mathbb{S})$ the folded string is that which takes the lowest value entropy. That means in a hypothetical hierarchy that the folded state should stay on the top of the hierarchy because it has the highest entropy loss with respect to the unfolded state. To search for a possible hierarchy we analyzed all the native substrings for the descriptions SRA[4] and SSS[8]. We took the native strings as reference, respectively "`0000210000000210000`" for SRA[4] and "`EEEESSEEEEEESSEEEE`" for SSS[4] and then decomposed in all the possible contiguos native substrings of length 1 to $R-1$. Implicitly we introduce a degree of nativeness by considering the number of native string sites in a generic string, imposing that the native substring must be contiguous. We call hierarchy level such degree of nativeness. We estimate from the simulation the probability of each substring.

The result of such calculations are the maps shown in figure 2.13. Plot (A) is referred to the SRA[4] while (B) to SSS[8]. In the diagonal the residue number corresponding to the string sites are reported, while the vertical axis is the hierarchy level. Colors in the map represent the probability of a native string fragment of length corresponding to the hierarchy level. For example the squares along the main diagonal which is made by $R = 18$ squares have hierarchy level 1 since there are $R = 18$ substrings of length equal to 1; the next diagonal has $R - 1 = 17$ squares and hierarchy level 2 as there $R - 1$ substrings of length 2 in a string of length R; and so on until reaching the full native string: hierarchy level $R = 18$. As mentioned squares are coloured according with their observed probability, thus the probability on the top of the hierarchy corresponds to that of the full native string. A strongly ordered heterogeneity appears on the maps which can be interpreted as the presence of patterns. Patterns can be organized

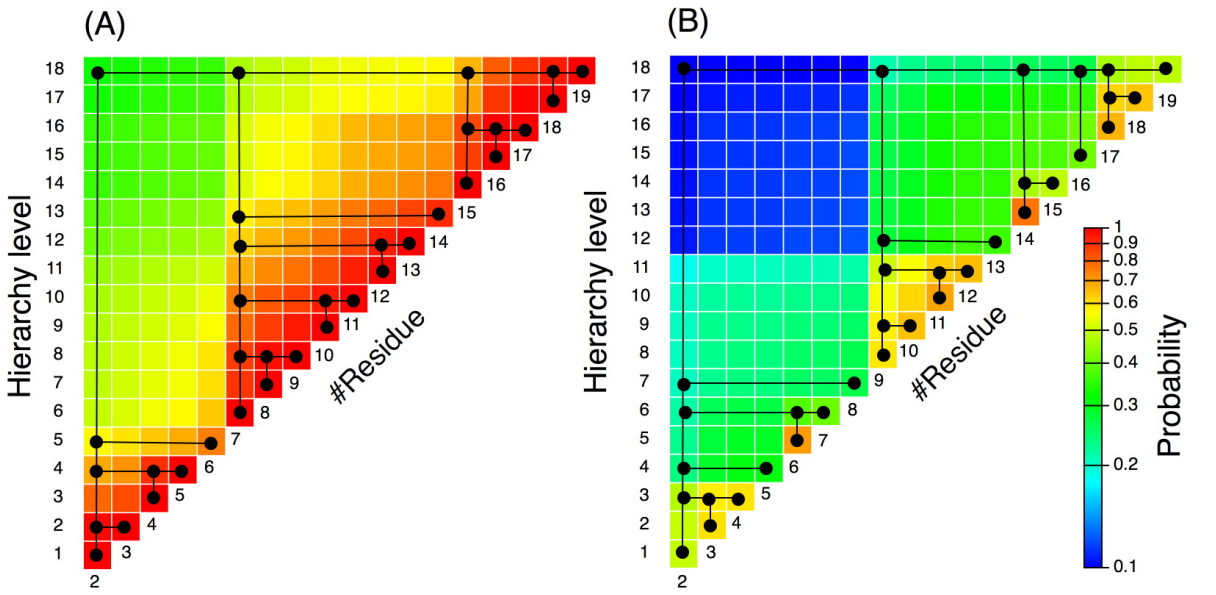

Figure 2.13: (A) SRA[4]; (B) SSS[8]. The configurational hierarchy of the folded string. Heterogeneity in the maps implies the existence of patterns in the configurational space. Patterns are revealed in terms of hierarchical trees which are constructed combining all the contiguous folded substrings in such a way to obtain fragments of the folded string that have maximal probability. For instance, at the lowest hierarchical level we have R 1-fragments which correspond to R nodes on the maps. At the next hierarchical level the 1-fragments are combined to obtain 2-fragments: two contiguous 1-fragments are combined if the probability of the corresponding 2-fragment is higher than the alternative 2-fragments. The 2-fragment so obtained gets a node which is linked to previous nodes of the 1-fragments, and so on. If a 1-fragment does not combine with any 2-fragment then it can be combined at next hierarchical level to form fragments of length longer than 2.

in a tree from the lowest hierarchy level till the highest. To show that an algorithm has been developed based on a weighted random walk. At the lowest hierarchical level R walkers start a walk, namely that 1-fragments must assembly in a certain way to gain the next status level of 2-fragments. The algorithm makes the walkers follow the maximal probability route, for instance a 1-fragments can assemble itself to two different 2-fragments, thus the algorithm choose that maximizing the fragment probability. The procedure is repeated for all the hierarchies until a tree is completed by reaching the full folded string. The algorithm finds the maximal probability tree associated to the map though in principle there may be other lower probability solutions, namely those corresponding to trees still satisfying the hierarchy. In figure 2.13 (A) the tree for SRA[4] shows a modular pattern: the first module from Trp2 to Gly6, the second module from Ser7 to Gly14 and a third module from Ser15 to Tyr19. Moreover, the second and third module are combined together at a high hierarchical level: longer fragments starting from that show a probability that is very close to that of the full folded string. This suggests that the combination of the module second and third (which corresponds to the second hairpin) might represent an intermediate state for folding, a state macroscopically well defined. Thus, in this description the modules appear to

be "parsed" by the positions of the glycines, which define the loops characterizing the triple stranded β-sheet folded state of the GSGS. The tree represents the most probable logical decomposition of the folded string which explains how the single residues should be combined in substrings to compose the folded state. Yet, a tree reflects the local character of the interactions that might be responsible of the folded state formation. Let's consider the panel (B) of figure 2.13, where the calculation is based on the SSS[8] description. In this case the modular pattern is more marked: glycine and serine residues result now combined at a low hierarchical level into a native-like turn configuration. The first module goes from Trp2 to Thr8, the second from Lys9 to Asn13 and the third from Gly14 to Tyr19. Also in this case the most probable tree have the second and third module combined at high hierarchy. However, in the second most probable tree (see in figure the white spots and the dashed links) it is the first module to be combined with the second. That means again that possibly two intermediate states are present: the most probable corresponding to the second hairpin formed, the second most probable corresponding to the first hairpin formed. The second intermediate state is more hidden in the description based on the rotational angles. This is another example where two descriptions lead to slightly different conclusions. One of the conclusions that can be drawn from the hierarchies investigated so far is that they seem to exclude an all-none cooperative folding mechanism typical for instance of two state proteins, at least for the system here investigated. The picture arising gives more importance to the relative role of the local interactions in determining the multistep and noncooperative character of the folding process, as already has been pointed out by Rose and colleagues [Baldwin and Rose, 1999a, Baldwin and Rose, 1999b, Rose et al., 2006]. That is particularly interesting despite the method here presented assumes, in first approximation, the residues interacting only contiguously or between first neighbor fragments. In principle, long range interactions in combining string fragments might play a relatively important role. Clearly such a generalized model would enormously increase the complexity of the hierarchy. An example of that can be found in the recent works of Dill [Hockenmaier et al., 2007, Ozkan et al., 2007, Voelz and Dill, 2007] in which the folding pathways are hierarchically constructed with zipping-assembly methods partially borrowed from modern linguistics. The limits of these quoted methods is that they are applied to lattice system, in which the protein chain is extremely simplified.

2.5.3 Convergence of the entropy

In this section we provide a method to judge whether a simulation reached convergence with respect to thermodynamic observables. We have already established that there is no convergence in the number N of sampled mesostates as they grow sub-linearly with the number of microstates M collected in a simulation. A natural quantity to be checked is the Shannon entropies associated with the ensemble of sampled mesostates, which are H and h for mesostates and strings respectively. From the viewpoint of information theory, these quantities evaluate the total amount of information that is stored in the whole ensemble of mesostate. That amount of information depends on the size of the system that we have seen to grow sub-linearly. Thus, once again it is convenient to consider directly the disorders $\mathcal{D}(H)$ and $\mathcal{D}(h)$ as defined in equations 2.34 and 2.65. In figure 2.14 the disorder production are shown as a function of the sampled microstates M, that are computed following and updating the mesostate and string site probabilities along the trajectories. Let's consider the function $\mathcal{D}(t) = H(t)/H_{max}(t)$ reported in figure 2.14 (A): we have $H_{max}(t) = k_B \ln N(t)$ with t the sampling time. If we take the derivative of $\mathcal{D}(t)$, we

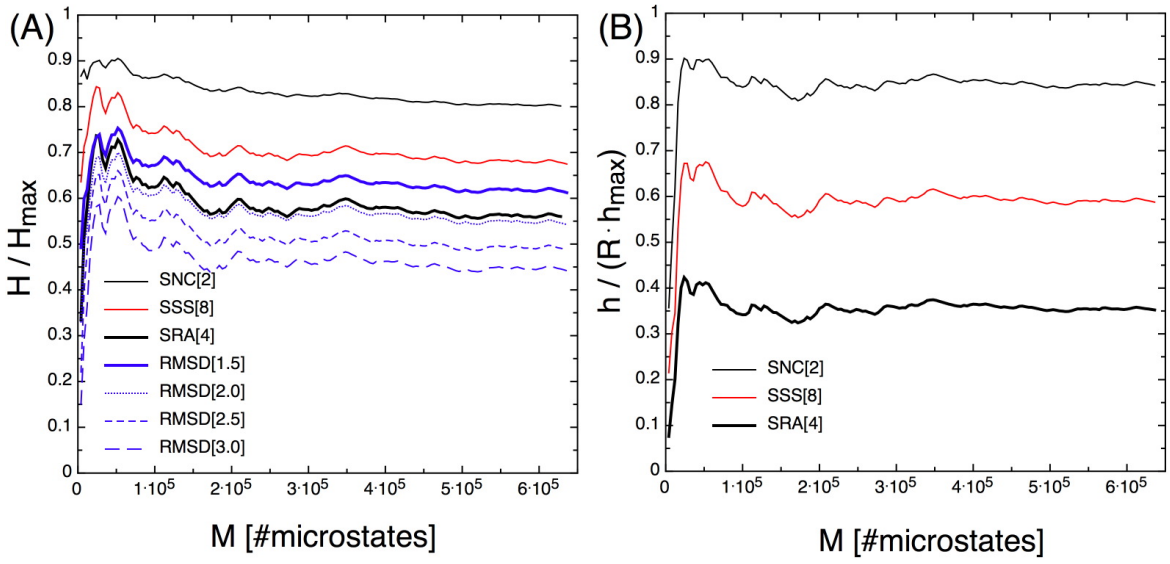

Figure 2.14: Disorder $\mathcal{D}(H)$ (A) and $\mathcal{D}(h)$ (B) as a function of the number of sampled microstates M. The convergence of the simulations is a signature of statistical convergence. This arises from the compensation between folding and unfolding events: former increase disorder while the latter decrease it.

easily obtain

$$\dot{\mathcal{D}}(t) = \left(\frac{\dot{H}(t)}{H(t)} - \frac{\dot{N}(t)}{N(t) \ln N(t)} \right) \mathcal{D}(t) \tag{2.66}$$

which means that the production rate of disorder is given by two terms with opposite sign, the first is the the rate of entropy production $\dot{H}(t)/H(t)$, while the second is the rate of mesostates production $\dot{N}(t)/N(t) \ln N(t)$. The sampling time t can be taken as $t = \tau M$ where τ is the lag time of the trajectory, which in our case is 20 ps. We have seen that the number of sampled mesostates N scales as M^{α} so that equation 2.66 can be rewritten as

$$\dot{\mathcal{D}}(M) = \left(\frac{\dot{H}(M)}{H(M)} - \frac{1}{M \ln M} \right) \mathcal{D}(M) \tag{2.67}$$

which does not depend on α and thus, consequently, does not depend on the description adopted, additionally, for large enough M it corresponds to the rate of entropy production. In the panel (A) of the figure we see that all the descriptions but SNC[2] reach a convergence value for the disorder when about one third of the microstates is sampled. The increase and decrease of the function corresponds to the folding events. In particular, when the system folds the disorder decreases as it visits parts of the conformation space already sampled, while when it unfolds the disorder increases because new regions of the configurational space are explored. The fact that a convergence is reached means that there is a compensation between folding and unfolding events: the decreasing of the disorder due to folding is compensated by its increasing due to unfolding. If one of the two processes overtakes the other one, then one obtain a monotonic disorder, namely

$$\frac{\dot{H}(M)}{H(M)} < 0$$

decreasing for folding and

$$\frac{\dot{H}(M)}{H(M)} > 0$$

increasing for unfolding. In figure 2.14 (B) we have the disorder calculated from the string entropy h, which shows qualitatively the same behaviour as in figure 2.14 (A). This result has a twofold importance: on one hand it demonstrates that the simulations are statistically representative of the dynamics of the system, on the other hand it explains how folded and unfolded states are closely related to each other. Moreover the convergence of the disorder can be interpreted in terms of convergence of the effective number of mesostates by taking the natural logarithm of the entropies H and h.

2.6 Folding kinetics in the space of mesostates

The estimation of rates, and in particular the rate for folding is particularly important to characterize the properties of the simulated model and possibly to compare its properties with experimentally measured kinetics. For our purposes it is also important to characterize the typical relaxation rates of the various mesostates making up the unfolded state. More in general, the characterization of the folding pathways connecting the unfolded mesostates to the folded mesostate using stochastic methods will be the main topic of this section.

In stochastic mechanics the rate of a reaction from a starting species A to a target species B is given by the reciprocal of the total escape time τ_e from A averaged over all the pathways connecting the two species. In the protein folding case, the determination of a rate is complicated by the difficulty of defining a variable that discriminates reactant and product. As we have seen so far, most coarse graining procedures essentially agree on the nature of the native state, although the size of the folded mesostate might depend of the type and the coarseness of the grain.

2.6.1 First passage times

In the following we only assume the existence of a coarse grained procedure that individuates a folded state. We have seen that all the reasonable coarse-graining procedures, despite the fact that they focus on considerably different properties, are able to individuate the native basin as the most populated. As we show in the following, the knowledge of the folded mesostate in terms of any coarse graining procedure is sufficient to robustly estimate a folding rate. Let us call X_F the most populated mesostate that corresponds to the folded mesostate and is the closest to the bottom of the free energy basin corresponding to the folded *macrostate*. The mesostate X_F defines a boundary between the very interior of the folded basin and all the other microstates that can be grouped in a generalized rest state X_R. Thus a mean first passage time t_{MFPT} to X_F can be estimated as

$$t_{\mathrm{MFPT}}(X_R \to X_F) = \int_0^\infty dt\, t F(X_R \to X_F; t) \tag{2.68}$$

where $F(X_R \to X_F; t)$ is the first passage time distribution (FPT). The function $F(X_R \to X_F; t)$ is the probability density that the FPT lies between t and $t + dt$, which is given by the conditional probability to have reached the target state X_F at the time t [Szabo et al., 1980, Hänggi and Talkner, 1985, Lee et al.,

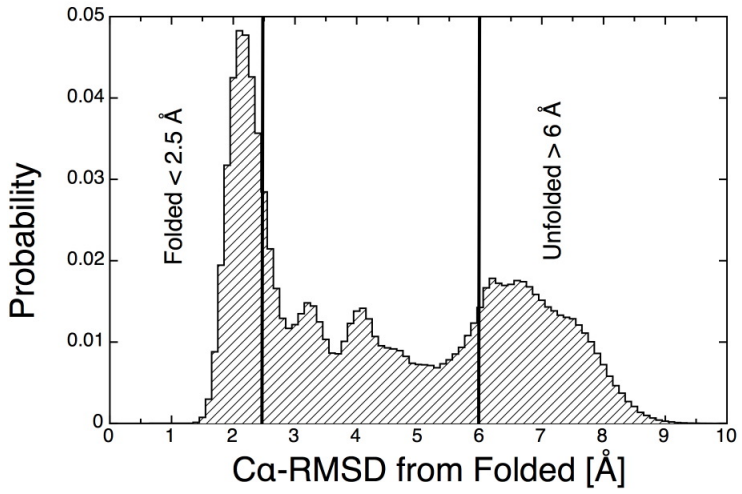

Figure 2.15: The Cα-RMSD distribution with respect to the folded structure of the GSGS from which a FPT distribution for unfolding is estimated. The unfolded target state has been chosen with Cα-RMSD greater than 6 Å.

2003] given the initial condition X_R,

$$\begin{aligned} F(X_R \to X_F; t) &= P(X_F|X_R; t) \\ &= 1 - P(X_R|X_R; t) \end{aligned} \tag{2.69}$$

where

$$P(X_F|X_R; t) = \frac{P(X_R \cap X_F; t)}{P(X_R)} \sim \frac{n(X_R \to X_F; t)}{n(X_R)} \tag{2.70}$$

with $P(X_R \cap X_F; t)$ is the joint probability for the transition $X_R \to X_F$ at the time t and $n(X_R \to X_F; t)$ is their total number given by

$$n(X_R \to X_F; t) = \sum_{t'} \theta_R(t')\theta_F(t' + t) \tag{2.71}$$

where the counter function $\theta(t)$ is defined in eq. (2.30). The target (folded) state X_F must satisfy the boundary condition

$$P(X_F|X_F; t) = 1 \tag{2.72}$$

To estimate the MFPTs from time series of mesostates we use the definition of conditional probability at time t so that the equation 2.68 becomes

$$\begin{aligned} t_{\mathrm{MFPT}}(X_R \to X_F) &= \int_0^\infty dt\, t P(X_F|X_R; t) \\ &\sim \frac{1}{n(X_R)} \sum_t \sum_{t'} t\, \theta_R(t')\theta_F(t' + t) \end{aligned} \tag{2.73}$$

which is easily computable from a time series. Thus the FPT distribution turns out as

$$F(X_R \to X_F; t) \sim \frac{1}{n(X_R)} \sum_{t'} \theta_R(t')\theta_F(t' + t) \tag{2.74}$$

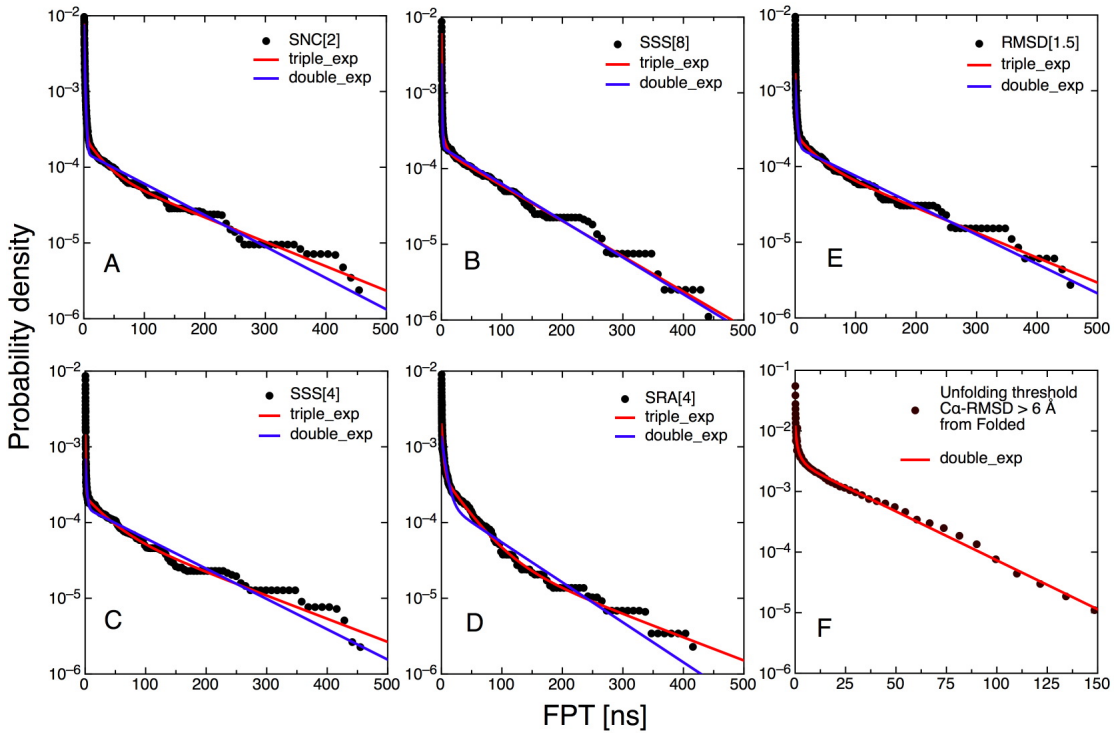

Figure 2.16: First passage time distributions to the folded mesostate for all the descriptions adopted to study the GSGS. Double and triple exponential functions were used to fit the data.

Generally if a target state is chosen at the boundary of the stochastic separatrix [Hänggi et al., 1990, Müller et al., 1997], i.e., the particular boundary from which the system either falls back to the initial state or directly reaches the final state with the same probability, the exchanging rate between the generalized mesostate X_R and the bottom of the folded state X_F is given by

$$k_f = \frac{1}{2t_{\mathrm{MFPT}}} = \frac{1}{\tau_f} \tag{2.75}$$

That is not the present case, as we have chosen the final state as the interior of the folded basin to avoid the problem of finding the stochastic separatrix. Nevertheless, the FPT distribution clearly shows a kinetic partitioning between the folded and the unfolded basins. In Figure 2.16 the FPT distributions for the four alternative coarse graining procedures are shown. In figure (F) the FPT distribution is also shown for unfolding events in which the target state has been chosen from the Cα-RMSD distribution to the folded structure greater than 6 Å (see figure 2.15). Different descriptions give slightly differences in the FPT distributions on the short time scales and substantial homogeneity in the middle-long time scale range. The differences on short and intermediate time scales are due to the fact that the four different coarse graining procedures somewhat differ. All the FPT distributions have been fitted with both triple and double exponential functions, whose slowest rate corresponds to a folding time that can be inferred from the fitting procedure. In Table 2.8 the reciprocal of the rates τ_d are shown for diffusion within the starting state of the FPT, τ_m for an intermediate phase (only for the triple exponential) and τ_f for folding. All the time scales are obtained by fitting the FPT distribution with either a triple or a double

Description	τ_d [ns]	θ_d	τ_m [ns]	θ_m	τ_f [ns]	θ_f	χ^2
SNC[2] triple-exp	1.2±0.2	$7 \cdot 10^{-3}$	25±19	$2 \cdot 10^{-4}$	134±28	$9 \cdot 10^{-5}$	5.1
SNC[2] double-exp	1.3±0.2	$7 \cdot 10^{-3}$	/	/	105±10	$1.5 \cdot 10^{-4}$	12.3
SSS[8] triple-exp	0.5±0.3	$5 \cdot 10^{-3}$	5±7	$1.7 \cdot 10^{-4}$	93±8	$1.7 \cdot 10^{-4}$	5.4
SSS[8] double-exp	0.9±0.3	$2.1 \cdot 10^{-3}$	/	/	89±6	$1.9 \cdot 10^{-4}$	7.1
SSS[4] triple-exp	0.7±0.5	$2 \cdot 10^{-3}$	34±25	$1.3 \cdot 10^{-4}$	142±36	$9 \cdot 10^{-5}$	3.7
SSS[4] double-exp	2±1	$5 \cdot 10^{-4}$	/	/	108±11	$1.5 \cdot 10^{-4}$	8.3
SRA[4] triple-exp	2.7±1.3	$1.5 \cdot 10^{-3}$	35±13	$4 \cdot 10^{-4}$	141±62	$5 \cdot 10^{-5}$	2.8
SRA[4] double-exp	6.5±1.8	$1.2 \cdot 10^{-3}$	/	/	82±10	$1.8 \cdot 10^{-4}$	12
RMSD[1.5] triple-exp	1.6±0.8	$1.4 \cdot 10^{-3}$	28±30	$1.4 \cdot 10^{-4}$	132±27	$1.3 \cdot 10^{-4}$	2.9
RMSD[1.5] double-exp	2.3±0.7	$1.2 \cdot 10^{-3}$	/	/	112±11	$1.8 \cdot 10^{-4}$	5.7
Unfolding Cα-RMSD> 6 Å	1.2±0.9	$8 \cdot 10^{-3}$	/	/	27±3	$3 \cdot 10^{-3}$	1.1

Table 2.8: Triple and double exponential fitting parameters of the FPT distributions for different descriptions of the configurational space.

exponential function

$$\begin{cases} F_{\text{triple}}(t) &= \theta_d e^{-t/\tau_d} + \theta_m e^{-t/\tau_m} + \theta_f e^{-t/\tau_f} \\ F_{\text{double}}(t) &= \theta_d e^{-t/\tau_d} + \theta_f e^{-t/\tau_f} \end{cases} \tag{2.76}$$

where the constants θ_d, θ_m and θ_f are amplitudes. Within the fitting errors the folding times obtained do not crucially depend on the used coarse graining procedures (except when a coarse graining based on RMSD and with a cutoff > 2 Å is adopted, data not shown) if the same method for fitting is adopted. The fitting results with a triple exponential show the presence of an intermediate phase with a relaxation time to the folded mesostate of about 25/35 ns for all the descriptions but the SSS[8]. In the latter case, the double and triple exponential fits essentially agree in giving a folding time of about 90 ns. For all the other descriptions a triple exponential better fits with the data than a double exponential. Double exponential exponential fit gives a folding time that is an average of the intermediate and folding time obtained with a triple exponential. The fast time scale represents a diffusion time of the microstates that are within the folded basin, and strongly depends on the kind of descriptors as it is related to the population of the folded mesostate. The unfolded basin can be kinetically detected as the ensemble of microstates having FPT approximatively $\gtrsim 50$ ns. As it will be clarified later, the unfolded ensemble relaxes to the folded basin with essentially single exponential kinetics. The intermediate time scale can be interpreted as a kinetic extension of the folded basin. Once the system get across the main folding free energy barrier it enters into a wide basin which is layered by several small free energy barriers. In other words, within this basin the system diffuses until the folded basin is reached with a mean time of about 25/35 ns. This intermediate phase can be called a "pre-folded" phase. This aspect is not clearly evident in the description SSS[8] due to redundancy of the secondary structure alphabet. In fact for SSS[4], where such a redundancy is suppressed (only helix, beta, turn, coil in the alphabet) the presence of an intermediate pre-folded phase is recovered. Yet, at this level of analysis, as we only selected a subset of the folded basin X_F as a target state for the FPT calculations, it is not surprising that we are not able to discriminate properties of the unfolded state, apart from an estimation of a lower limit of the folding

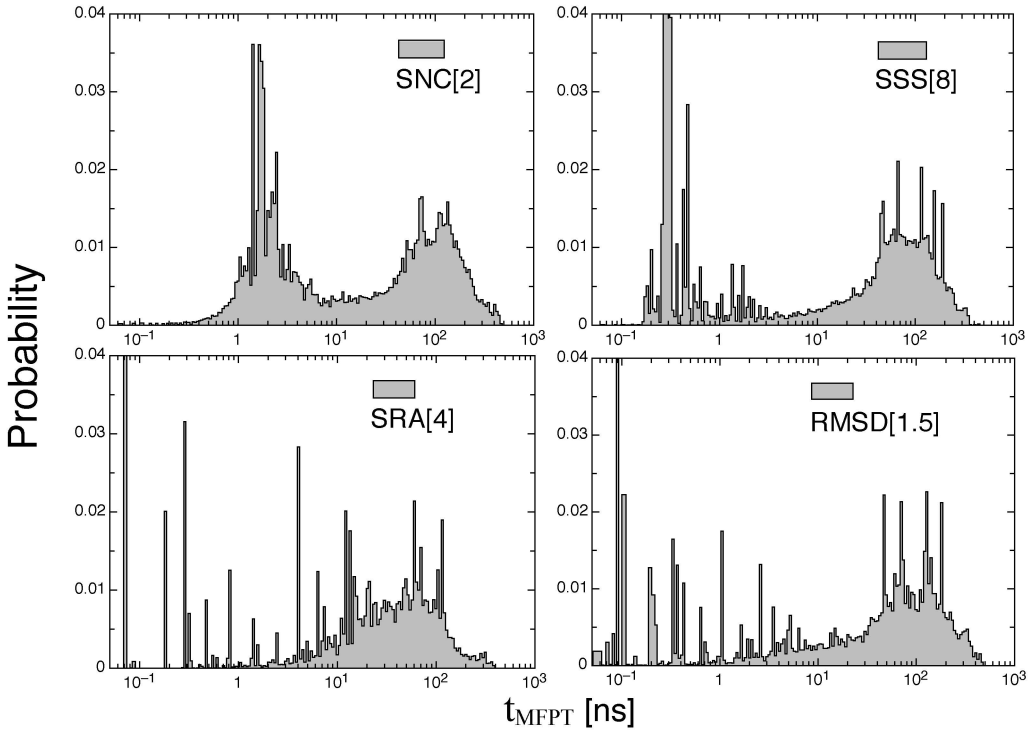

Figure 2.17: The mean first passage time MFPT distributions to the folded mesostate for the adopted descriptions. The distributions show a clear kinetic partitioning at all time scales in both the folded and unfolded phases: the pronounced peaks in the main unfolded region correspond to mesostates having well defined relaxation times to the folded state.

time. The unfolding FPT distribution was fitted with a double exponential and the estimated unfolding time inferred with the boundary on the Cα-RMSD results of about 27 ns which is quite similar to the relaxation time from the intermediate pre-folded phase to the folded suggesting that the main reaction to unfold corresponds to a diffusion in the pre-folded phase.

2.6.2 Mean first passage times

For a more detailed study of the kinetics, one can use the whole set of mesostates obtained for all four different coarse graining procedures used in this work. In this way no assumption is made on the boundary of the folded state. However, only the most populated macrostate is assumed to be the target state of all the folding pathways starting from all the possible mesostates. If the system is partitioned in a set of mesostates $X_1, ..., X_N$ ranked by decreasing probabilities P_i, the state X_1 corresponds to the bottom of folded free energy basin X_F as in the previous calculations. Given the couple of states X_i ($i > 1$) and $X_F = X_1$ the MFPT $t_{\mathrm{MFPT}}(X_i \to X_F)$ for the reaction $X_i \to X_F$ is given by

$$t_{\mathrm{MFPT}}(X_i \to X_F) = \int_0^\infty dt\, tF(X_i \to X_F; t) \tag{2.77}$$

where $F(X_i \rightarrow X_F; t)$ is the first passage time (FPT) distribution for the reaction $X_i \rightarrow X_F$. Given the initial condition X_i

$$\begin{aligned} F(X_i \rightarrow X_F; t) &= P(X_F|X_i; t) \\ &= 1 - \sum_{j \neq F} P(X_j|X_i; t) \end{aligned} \tag{2.78}$$

when $P(X_F|X_i; t)$ is the conditional probability at time t

$$P(X_F|X_i; t) = \frac{P(X_i \cap X_F; t)}{P(X_F)} \sim \frac{n(X_i \rightarrow X_F; t)}{n(X_F)} \tag{2.79}$$

with $P(X_i \cap X_F; t)$ the joint probability for the transition $X_i \rightarrow X_F$ at the time t and $n(X_i \rightarrow X_F; t)$ is the total number of these given again by

$$n(X_i \rightarrow X_F; t) = \sum_{t'} \theta_i(t')\theta_F(t' + t) \tag{2.80}$$

With the boundary condition $P(X_F|X_F; t) = 1$ on the target state X_F we obtain

$$\begin{aligned} t_{\mathrm{MFPT}}(X_i \rightarrow X_F) &= \int_0^\infty dt\, tP(X_F|X_i; t) \\ &\sim \frac{1}{n(X_i)} \sum_t \sum_{t'} t\, \theta_i(t')\theta_F(t' + t) \end{aligned} \tag{2.81}$$

All the MFPTs to the X_F state are computed for each the mesostates X_i . The calculations are carried out for the four different coarse graining procedures considered. The distributions of MFPTs have been estimated and they are shown in figure 2.17 for the descriptions SNC[2], SSS[8], SRA[4], RMSD[1.5]. The differences of these distributions from those of FPT (figure 2.16), having divided the conformation space by two macrostates X_F and X_R are immediately evident. In the former the exponential character of the folding process appears clearly. In the MFPT distributions the state X_R is partitioned in multiple mesostates resulting in a higher complexity of the folding kinetics so that the exponential character appears to be broken. In particular, both the folded and the unfolded basins appear to be layered into different time scales. According to the SNC description, folded and unfolded basins seem kinetically partitioned while in the other descriptions this distinction is less clear. What is clear, on the other hand, is that the mesoscopic configurational space has multiple minima, which are kinetically well defined. Thus what appears exponential in the microscopic FPT distribution, due to the self averaging in the phase space, at mesoscopic level is indeed differentiated: it follows that the mesoscopic configurational space is structured.

The MFPT from all the mesostates to the folded one can be thought as reaction coordinate that gives how kinetically far from the folded basin a mesostate is. On this respect it is useful to express the thermodynamics in terms of the MFPT which is a kinetic quantity. We have thus computed, on log-binned values of MFPTs, the mean effective energy (potential plus solvation energies) and the mean entropy S. The entropy S is calculated using the equation 2.24, $S = h + S_{\mathrm{b}}$ where h is the string configurational entropy defined by 2.62. All the values are referred to the shortest time scale, which turns out to be a Δ value with respect to the folded state. Having the ΔE_F and ΔS_F we can calculate the free energy difference as $\Delta G_F = \Delta E_F - T\Delta S_F$. The result of the calculations for the description SRA[4] based on the rotational angles are shown in figure 2.18. The plot is paradigmatic; the free energy profile

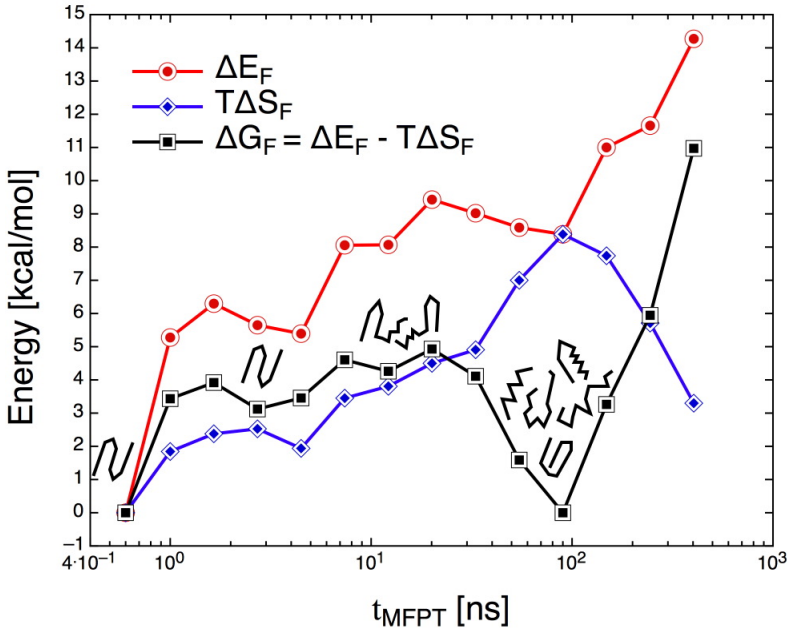

Figure 2.18: The relation between thermodynamics and kinetics by using the MFPT as a reaction coordinate and SRA[4] as coarse graining.

shows two main minima at $t_{\mathrm{MFPT}} < 1$ ns and $t_{\mathrm{MFPT}} \sim 100$ ns, which correspond to the folded and the unfolded basins respectively. Interestingly, the unfolded minimum for the free energy is characterized by a maximum of the entropy. In particular, the effective energy and the free energy contribution due to the entropy compensate each other giving a $\Delta G_F \sim 0$, which correctly reflects that the simulation temperature of 330 K is that of melting. Furthermore, the free energy profile shows also other two shallow intermediate minima, one at about 20 ns, the other at about 4 ns. The first corresponds to peptide configurations in which one of the two harpins are formed and the other is unstructured, while the second adopts the topology of the folded state, but more "floppy". Figure 2.18 shows that it is possible to use as unidimensional reaction coordinate upon the condition that the employed coarse graining procedure is reasonable enough to preserve the relevant information of the process kinetics. The MFPTs of selected mesostates are reported in table 2.9, the selection corresponds to that of table 2.7 and figure 2.9. From the values reported, it follows that the unfolded free energy basin is quite heterogeneous and mainly populated by a helical basins.

2.6.3 Folding kinetics hierarchy

In section 2.5.2 the hierarchy of the native substrings has been analyzed by estimating all the probabilities for the native substrings having length from 1 to $R-1$, with R the full string length. It was clear that under a probabilistic point of view a hierarchy of configurational states naturally arises, essentially determined by the closure of the 1st and 2nd beta hairpins to form the three stranded beta sheet. In this section we want to study the hierarchy assembly from a kinetic point of view. Thus, instead of estimating the probabilities, the MFPTs of all possible native substrings are computed, starting from the MFPT of a single native string site to the full native string. Again, we have taken the native

Id number	SRA string	t_{MFPT} [ns]	P_i [%]
1	**000021000000210000**	**0.06**	**32.3**
5	000011000000210000	6.5	2.1
7	010011000000210000	113	1.2
8	000020000000210010	40	0.9
9	100021000000210000	1.2	0.8
10	000020010000210000	29	0.7
15	000020000000210000	17.3	0.5
25	000021000000210011	8.5	0.2
27	010001000000210000	104	0.2
33	000021000000100000	23	0.2
40	010001111111111111	69	0.1
42	010010000000210000	34	0.1
47	001020010000210000	33	0.1
65	000111001000210010	108	0.08
70	111111111111300000	123	0.07
82	100021000100110000	13	0.06
89	010001111111300000	102	0.06
127	010001111111111100	80	0.04
136	000020010000110000	31	0.04
153	111111111111111111	88	0.03

Table 2.9: The MFPTs of selected mesostates to the folded mesostate taken from table 2.7 and figure 2.9.

strings as reference, respectively "000021000000210000" for SRA[4] and "EEEESSEEEEEESSEEEE" for SSS[4] and then decomposed in all the possible contiguos native substrings of length 1 to $R-1$. For each native substring we estimate the MFPT necessary to reach it. As a result we obtain the maps of figure 2.19, where plot (A) is referred to the SRA[4] while (B) to SSS[8]. In diagonal the residue number corresponds to the string sites, while in the vertical axes the hierarchy level. Color shades in the map represent the reciprocal of the MFPT (the rate) to the native string fragment of length corresponding to the hierarchy level. To reconstruct the possible hierarchies of the maps the same weighted random walk algorithm applied in section 2.5.2 is used. In the previous case the algorithm finds the maximal probability tree, while in the present case, the result is the maximal rate tree shown by the black nodes and links of figure 2.19. In other words, R walkers starting from the lowest hierarchy level assemble into each other to form higher hierarchy fragments until the full folded string is obtained. To give an example, two contiguous 1-fragments are assembled together to form a 2-fragment if the formation rate of that is the highest possible among all the possible 2-fragments that contiguous 1-fragments can form. In particular, nodes and links (depicted in black in figure 2.19) show how the folded substrings (nodes) are assembled together (links) to form the full folded string in a fastest way. Again, a modular structure of the trees is found, although they qualitatively differ from those obtained with a probabilistic criteria. Previously the positions of the Glycines played the role of module breakers, while in the present case the tree structure appears to be complementary to that. The residues responsible of the beta turns are the fastest to reach their folded state (about 1 ns) so that, the propagation of the folding signal initiates and diffuse from them. For the SRA[4] description the maximal rate tree is shaped by 2 modules: a N-term module from Trp2 to Trp10 and a C-term module from Tyr11 to Tyr19. The two modules are organized around the formation of the beta turns. The modules can be interpreted as two main folding pathways,

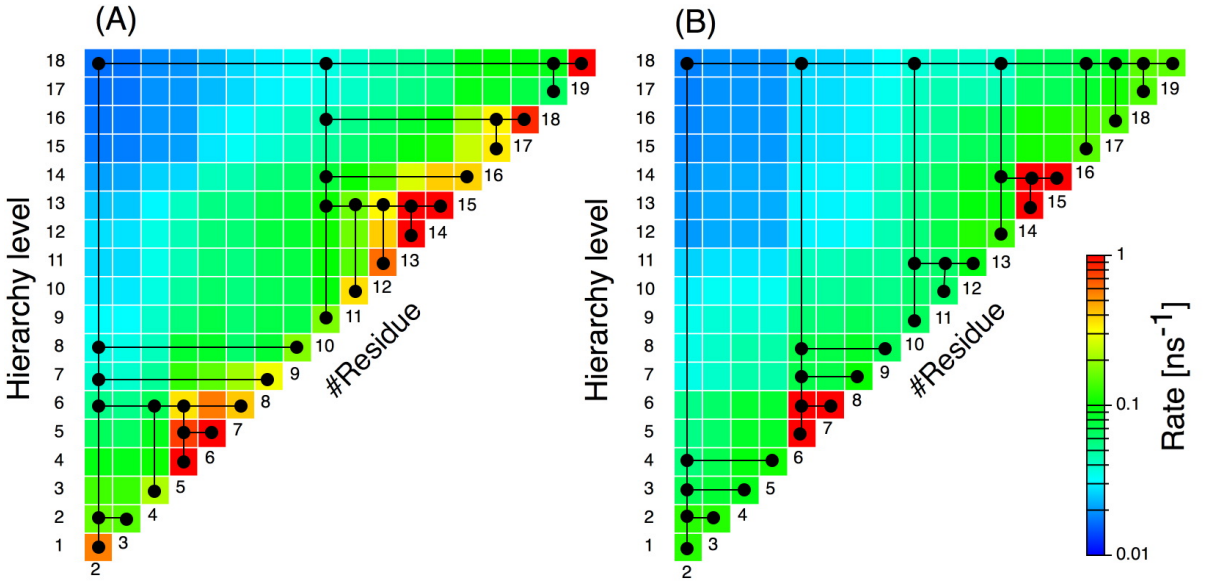

Figure 2.19: The kinetic hierarchies of the folded string for the descriptions SRA[4] (A) and SSS[8] (B). A pattern of folding pathways appears in figure 2.13 as well. The trees composed by black nodes and edges are extrapolated from the maps by using the same algorithm employed to construct the trees of figure 2.13. For instance, two contiguous 1-fragments (two nodes at the lowest hierarchy) are assembled together to form a 2-fragment (giving a new node at the next hierarchical level) if the formation rate of it is the highest possible among all the possible 2-fragments that 1-fragments can form. The trees show the fastest way in which the folded string can be assembled from all folded substrings.

where one of the two hairpins is formed first. The pathway in which the second hairpin is closed first is slightly more favorable than the other, as its formation rate is an order of magnitude higher (compare the fragments from residue 2 to 9 with rate ~ 0.04 ns^{-1} and from residue 10 to 18 with rate ~ 0.1 ns^{-1}). In the panel (B) of figure 2.19 the hierarchical tree is more branched. In particular, the folding process in this case appears serial: first the second hairpin is formed, then the first. In this description folding is then a serial sequence of events rather than parallel as in the other description. The trees shown here correspond to the maximal rate tree of the weighted random walk, namely the fastest way to assemble all possible folded substrings, other "lower" rate trees with parallel pathways are also possible.
The modular structure of the maps is anyway less structured in the present case than in that based on the fragment probabilities (figure 2.13). This suggests that under the hypothesis made, if studied from a kinetic viewpoint, folding process appears smother than in the complementary thermodynamic description: it is more cooperative. Therefore, though the highest rate folding pathway is that in which the second beta hairpin closes first, a multiplicity of other pathways cannot be excluded solely on the basis of these results. What appears again clear from this kind of analysis, is how the results and the interpretations on folding mechanisms, strongly depend on the adopted descriptors. Thus, before drawing final conclusions on such a complex phenomena, it is a good habit to collect as much as possible of diverse data.

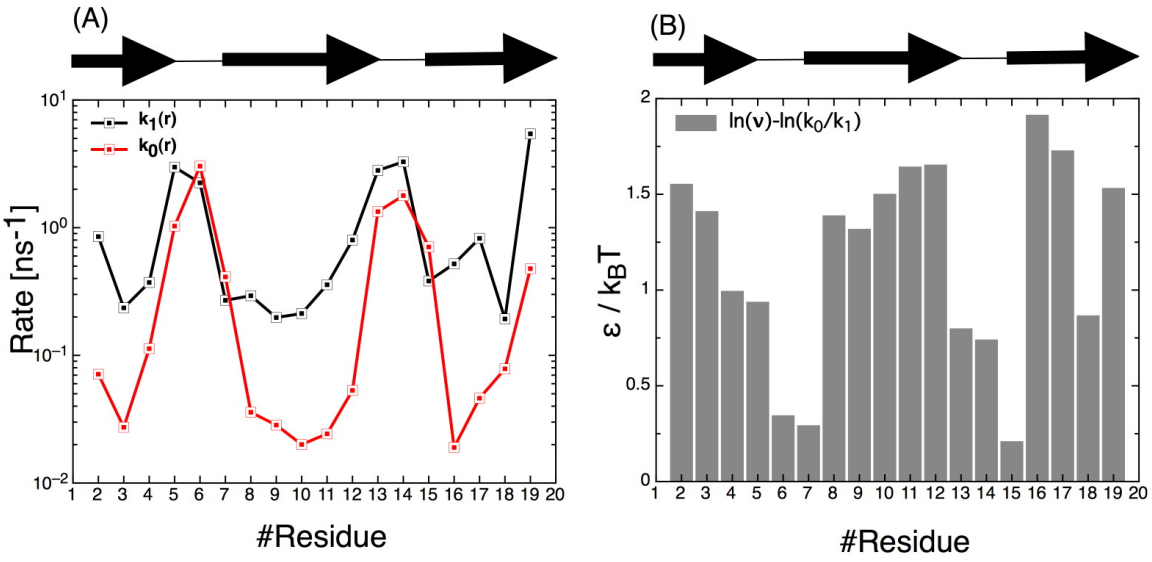

Figure 2.20: (A) The local rates for the mesoscopic transitions k_0 ($c \to i$) and k_1 ($i \to c$) that are used for the generalized Zwanzig model according to the SRA[4] description. (B) The ratio ϵ/k_BT corresponding to the favorable energy bias to a correct bond chain from equation 2.83.

2.6.4 Local rates and the Zwanzig model

In the previous section we investigated the kinetics of forming a native string starting from the combinations of all possible substrings of the native string. In this section we take a more minimalist approach to give account of the folding of the GSGS. We inspire to the simple model introduced by Zwanzig [Zwanzig et al., 1992, Zwanzig, 1995, Zwanzig, 1997] to discuss the Levinthal paradox [Levinthal, 1968]. Levinthal's paradox is that, searching the native folded state of a protein by a random search among all possible configurations, can take a non biological long time. In the modern illustration of the Levinthal paradox, each bond connecting amino acids can have several (typically three) states so that a protein of 100 amino acids could exist in $3^{100} \sim 5\times10^{47}$ configurations. If to sample a configuration takes about a fs ($\sim 10^{-12}$ s) then to cover the whole configurational space it would take about 10^{27} years, a time several orders of magnitude greater than the age of the universe. The paradox consists in the fact that proteins fold nevertheless. Zwanzig in discussing it, wanted to show that even without the need of the energy landscape theories, one can obtain biological folding times by virtue of introducing a local favorable bias to the folded state. His model is an Ising-like one dimensional model (for a review of such models in protein folding see [Muñoz, 2001] and references therein) in which the protein is a chain of $R+1$ amino acids with R bonds that can be in either a correct "c" or incorrect "i" state, where correct means native and incorrect is non native. Quoting Zwanzig: *starting with an arbitrary distribution of correct and incorrect bonds, and some rule for making changes, find how long it takes to get to the perfect chain for the first time.* As a changing rule of the bond state, pseudo microscopic (mesoscopic) exchange rates are introduced: k_0 for c→i and k_1 for i→c. Consequently, the number s of incorrect bonds in the protein configurations change along the time. Thus, the FPT to the full correct state is the necessary time, starting from any initial s, to arrive for the first time at $s = 0$. The MFPT $\tau(s)$ is the mean value of the FPT times from s to $s = 0$. In

case of identical bonds Zwanzig found

$$\tau(1) \cong (1/Rk_0)(1 + k_0/k_1)^R \tag{2.82}$$

whose result crucially depends on the ratio k_0/k_1. This ratio is the equilibrium constant

$$K = k_0/k_1 = \nu e^{-\epsilon/k_B T} \tag{2.83}$$

of a chain site, with ν the degeneracy of the incorrect state and ϵ the energy difference of the two states. One can obtain a biological time scale of 1 sec. already for an energy bias of $2k_BT$ in favor of the correct state, meaning that cooperative folding *is not* a necessary prerequisite to achieve short time scales for folding. In the original Zwanzig model all the bonds of the protein chain are indistinguishable, so that there are unique k_0 and k_1 rates. In this section we consider a generalized model in which the protein chain can have different bonds with different rates estimated from the simulations. The descriptor we choose for the present treatment is that based on torsional angles SRA[4]. We take the string of the native state as full correct folded mesostate `000021000000210000`. For each of the string site $r = 1, ..., R$ we estimate, from the simulations, the rates $k_1(r) = 1/t_{\mathrm{MFPT}}(i \to c; r)$ and $k_0(r) = 1/t_{\mathrm{MFPT}}(c \to i; r)$, where $t_{\mathrm{MFPT}}(r)$ are MFPT times calculated with the methods previously introduced. In figure 2.20 (A) the corresponding rates of the GSGS based on the SRA[4] string description are shown: for the beta portions of the folded state the rates k_1 are rather faster than the k_0. For the sites involved in beta turns the formation-distruption rates are comparable though the second beta turn forms faster, supporting again the conclusion that the pathway where the second harpin is formed first is more populated. In (B) we show the energy biases ϵ/k_BT to the correct sites, all the biases are more favorable to the correct states than to the incorrect. Let us consider then a chain with s incorrect bonds with the changing rates of figure 2.20 (A). We have two main rates, that for the process like $s \to s + 1$

$$\mathrm{rate}(s \to s+1) = K_+(R, s) = \sum_{\mathrm{R-s\ correct\ sites}} k_0(r) \tag{2.84}$$

and that for the process like $s \to s - 1$

$$\mathrm{rate}(s \to s-1) = K_-(R, s) = \sum_{\mathrm{s\ incorrect\ sites}} k_1(r) \tag{2.85}$$

Using combinatorics one can easily see that the two rates turns out to be

$$K_+(R, s) = \binom{R-1}{R-s} \sum_{r=1}^{R} k_0(r) = \binom{R-1}{R-s} K_0,\ s > 0 \tag{2.86}$$

and

$$K_-(R, s) = \binom{R-1}{s} \sum_{r=1}^{R} k_1(r) = \binom{R-1}{s} K_1,\ s > 0 \tag{2.87}$$

Having understood what are the main rates of the process, one can construct the corresponding master equation with respect the variable s for the probability $P(s, t)$ to have a string with s incorrect sites, as shown in figure 2.21. The main assumption is that only contiguous transitions in the variable s are possible. This leads to the equation

$$\begin{aligned} \frac{d}{dt}P(s,t) &= K_+(R, s-1)P(s-1,t) + K_-(R, s+1)P(s+1,t) - \\ &- K_+(R, s)P(s,t) - K_-(R, s)P(s,t) \end{aligned} \tag{2.88}$$

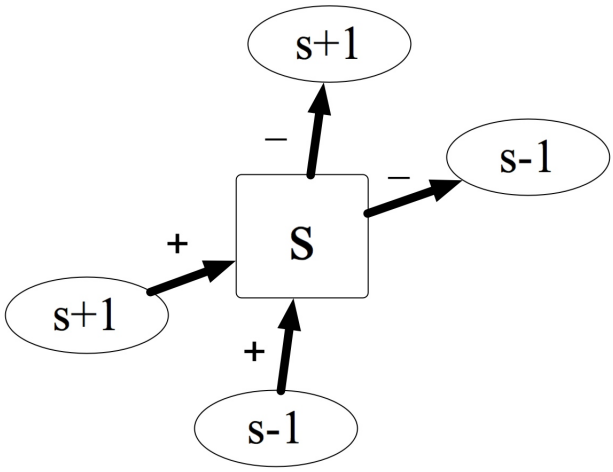

Figure 2.21: The procedure to construct the master equation on the s quantity.

that can be written in a matrix form

$$\frac{d}{dt}P(s,t) = \sum_{s'=0}^{R} W(s,s')P(s',t) \tag{2.89}$$

Equation 2.88 corresponds to a "birth-and-death process" or a "one-step process" [van Kampen, 1981]. The rate matrix $W(s,s')$ has the following structure

$$W = \begin{pmatrix} -K_+(R,0) & K_+(R,0) & 0 & \cdots & \cdots & 0 \\ K_-(R,1) & -(K_-(R,1)+K_+(R,1)) & K_+(R,1) & 0 & \cdots & 0 \\ 0 & K_-(R,2) & -(K_-(R,2)+K_+(R,2)) & K_+(R,1) & 0 & 0 \\ & & \vdots & \vdots & \ddots & \\ 0 & 0 & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & K_-(R,R) & -K_-(R,R) \end{pmatrix} \tag{2.90}$$

Instead of calculating the MFPT problem with the adsorbing boundary for $s = 0$ as in [Zwanzig et al., 1992] and [Szabo et al., 1980] we prefer to study the eigenvalues spectra of the rate matrix W, as also applied in [Henry and Eaton, 2004]. All the characteristic relaxation rates of the system are provided by the eigenvalues spectra of the rate matrix, and for a two-state protein the overall relaxation rate may be estimated from the smallest non-zero eigenvalue. For the matrix W constructed with the data collected we have $K_0 = 9.3$ ns^{-1} and $K_1 = 22.1$ ns^{-1} that with $R = 18$ gives the smallest non zero eigenvalue $k_{overall} = 0.035$ ns^{-1} that gives a time scale of about 30 ns. This time scale is shorter than the reciprocal of the folding rate we have previously obtained, about 100 ns, but it has nevertheless about the same order of magnitude. Despite highly approximated, the result appears to be interesting as it implies that a realistic folding time can arise from an intrinsic non-cooperative process, where only local biases towards the folded state play a leading role.

2.7 Causal mesostates and conformational master equation

In the last five years many authors approached the protein folding problem and the computer simulations of bio-molecules essentially as a *data mining* issue [de Groot et al., 2001, Nerukh et al., 2004, Swope

et al., 2004a, Lenz et al., 2004, Park and Pande, 2006, Noe et al., 2007, Chodera et al., 2007]. Purpose of data mining is the development of computational algorithms for the identification, or the extraction, of patterns from complicated data. Aims are essentially the data analysis of complex processes that are characterized by a high dimensionality and chaotic behavior. Essentially the idea is: given a process probed with respect to a set observables as a function of the time, one wants to construct a *computational machine* that is able to: i) reproducing the main phenomenology of the original observed process, ii) predicting/discovering previously unknown features about the process. Focusing on the use of computer simulations in protein folding, one would like to fit a minimal model on the simulation data and then using it to describe the relevant mechanisms of folding. Typically, the simplest models are based on Markov processes. Markovian processes, given a set of dynamic states allow to temporally evolve a system without taking into account its past dynamic history. Among the first works that pioneered predictive Markov models in protein folding, there are the works of Cieplak and Dill [Cieplak et al., 1998, Ozkan et al., 2001, Ozkan et al., 2003] where the time evolution of 2-dimensional Go-based lattice models have been studied using a Markovian master equation. These studies have given an important insight in the understandings both folding kinetics and also the physical meaning of Φ-values [Ozkan et al., 2003]. A crucial issue of the phenomenological Markov models is to verify that the underlying dynamics is intrinsically markovian [Park and Pande, 2006], a property whose verification generally deeply depends on the definition of adopted observables [Swope et al., 2004a, Swope et al., 2004b]. If the original simulations are MD simulations, long memory can in general characterize to relaxation dynamics of the degrees of freedom of the system [Nerukh et al., 2004]. In the following we use our coarse grained description of the GSGS from MD simulations to construct a markovian computational machine whose intent is to describe the kinetics of folding. Paraphrasing van Kampen [van Kampen, 1981], finding observables which make a process markovian is truly the art (duty) of the theorist.

2.7.1 Markov approximation

A Markov process is a stochastic process in which a certain dynamic variable x has the property that for any set of n consecutive times ($t_1 < t_2 < \cdots < t_n$) one has

$$T_{1|n-1}(x_n, t_n|x_1, t_1; \cdots; x_{n-1}, t_{n-1}) = T_{1|1}(x_n, t_n; x_{n-1}, t_{n-1}) \tag{2.91}$$

which means that the conditional probability density at t_n, given the value x_{n-1} at t_{n-1} is only determined and not affected by any knowledge of the past time values [van Kampen, 1981]. $T_{1|1}$ is the so called transition probability. The configurational space of a polypeptide chain is coarse grained in a set of mesostates that are used to classify an ensemble of sampled microstates by means of MD simulation. Each type of configurational description admits an upper limit N in the number of accessible mesostates. Indicating with ω_i a mesostate ($i = 1, ..., N$) we call $P_i(t)$ the occupation probability at the time t of the mesostate ω_i. Assuming that the system is closed and isolated, the jumps between mesostates ω_i can be described as a Markovian jumping process such that the occupation probability at the time t $P_i(t)$ is governed by the mesoscopic master equation

$$\frac{d}{dt}P_i(t) = \sum_{j=1}^{N} (T_{ij}(\tau)P_j(t) - T_{ji}(\tau)P_i(t)) \tag{2.92}$$

given the initial condition $P_i(0)$. The matrix $T_{ij}(\tau) = T(\omega_j|\omega_i;\tau)$ is the $N \times N$ transition rate matrix for the transitions $\omega_i \rightarrow \omega_j$ at the lag time τ. To the equation 2.92 corrisponds a steady state distribution P_i^e that is the equilibrium distribution of the system such that $\sum_{i=1}^{N} P_i^e = 1$. The thermodynamical interpretation of the steady state distribution follows directly introducing the conformational entropy H_i of mesostate

$$H_i = -k_B T \ln P_i^e \tag{2.93}$$

and a total conformational entropy

$$H = -k_B T \sum_{i=1}^{N} P_i^e \ln P_i^e \tag{2.94}$$

The transition rate matrix $T_{ij}(\tau)$ has a kinetical interpretation providing the mesoscopic free energy barriers such that

$$\Delta G_{ij}^{\ddagger}(\tau) = -k_B T \ln T_{ij}(\tau) \tag{2.95}$$

In general the matrix $T_{ij}(\tau)$ is not simmetric and the relation between opposite transition probabilities should satisfies the detailed balance condition for systems at their thermal equilibrium

$$P_i^e T_{ij}(\tau) \sim P_j^e T_{ji}(\tau) \tag{2.96}$$

which tells that at the equilibrium, the forward and backward probability fluxes should compensate each other between two well defined mesoscopic states. The mesoscopic master equation (2.92) is completely true only theoretically. Under a phenomenological viewpoint we consider the data given by an equilibrium MD simulation in the context of a Markov process. The finiteness of the sampling (in time and space) of the MD allows an approximated markovian description of the dynamics, if and only if the underlying process is at some extent Markovian. Later a test of the markovianity of the process is proposed, before that we introduce the formalism to estimate the transition rate matrix from simulation data.

2.7.2 Estimating a stochastic matrix

We want to estimate the matrix $T_{ij}(\tau)$ from a finite MD trajectory of microstates where the time window between microstates is discrete and where all the microstates are mapped in mesostates. The transition rate $T_{ij}(\tau)$ is the 1-step conditional probability for time unit τ to jump to ω_j, standing at the state ω_i at the previous step, namely

$$T_{ij}(\tau) = \frac{P_{ij}^e(\tau)}{P_i^e} \tag{2.97}$$

where $P_{ij}^e(\tau) = P^e(\omega_i \cap \omega_j;\tau)$ is the corrisponding 2-point joint probability for that transition or its total flux. The 2-point joint probability is estimated as time average over all the 2-point transitions along the trajectory

$$\begin{aligned} P_{ij}^e(\tau) &= \frac{1}{M-1}\sum_t \theta_i(\omega(t))\theta_j(\omega(t+\tau)) \\ &= \langle \theta_i(t)\theta_j(t+\tau)\rangle_t \end{aligned} \tag{2.98}$$

where M is the total length of the simulation and the θ function is the counter function defined in 2.30. It holds the following normalization

$$\begin{cases} \sum_{j=1}^{N} T_{ij}(\tau) = 1 \\ \sum_{i=1}^{N}\sum_{j=1}^{N} P^e_{ij}(\tau) = 1 \end{cases} \tag{2.99}$$

The asymptotic properties in the time scale τ of the 1-step transition rate probabilities derive from the asymptotic properties of the 2-point joint probabilities:

$$\begin{cases} \lim_{\tau\to 0} P^e_{ij}(\tau) = \langle \theta_i\theta_j \rangle_t = P^e_i \delta_{ij} \\ \lim_{\tau\to\infty} P^e_{ij}(\tau) = \langle \theta_i \rangle_t \langle \theta_j \rangle_t = P^e_i P^e_j \end{cases} \Rightarrow \begin{cases} \lim_{\tau\to 0} T_{ij}(\tau) &=& \delta_{ij} \\ \lim_{\tau\to\infty} T_{ij}(\tau) &=& P^e_j \end{cases} \tag{2.100}$$

where δ_{ij} represents the Kroeneker symbol.

To investigate the Markov approximation at a given time scale τ it is necessary to define a 2-step conditional probability involving three mesostates for the transitions $\omega_i \to \omega_j \to \omega_k$, such as

$$T_{ijk}(\tau) = \frac{P^e_{ijk}(\tau)}{P^e_{ij}(\tau)} \tag{2.101}$$

where $P^e_{ijk}(\tau) = P^e(\omega_i \cap \omega_j \cap \omega_k; \tau)$ is a 3-point joint probability which is still obtained as a time average over the whole trajectory, namely

$$\begin{aligned} P^e_{ijk}(\tau) &= \frac{1}{M-2}\sum_t \theta_i(\omega(t))\theta_j(\omega(t+\tau))\theta_k(\omega(t+2\tau)) \\ &= \langle \theta_i(t)\theta_j(t+\tau)\theta_k(t+2\tau) \rangle_t \end{aligned} \tag{2.102}$$

Again, we have the normalization

$$\begin{cases} \sum_{k=1}^{N} T_{ijk}(\tau) = 1 \\ \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} P^e_{ijk}(\tau) = 1 \end{cases} \tag{2.103}$$

The asymptotic properties of the 3-point joint probability derive from the asymptotic properties of the 2-step conditional probability $T_{ijk}(\tau)$:

$$\begin{cases} \lim_{\tau\to 0} P^e_{ijk}(\tau) = \langle \theta_i\theta_j\theta_k \rangle_t = P^e_i \delta_{ij}\delta_{jk} \\ \lim_{\tau\to\infty} P^e_{ijk}(\tau) = \langle \theta_i \rangle_t \langle \theta_j \rangle_t \langle \theta_k \rangle_t = P^e_i P^e_j P^e_k \end{cases} \Rightarrow \begin{cases} \lim_{\tau\to 0} T_{ijk}(\tau) &=& \delta_{jk} \\ \lim_{\tau\to\infty} T_{ijk}(\tau) &=& P^e_k \end{cases} \tag{2.104}$$

The evaluation of the 2-step conditional probability $T_{ijk}(\tau)$ is important to quantify how much the mesoscopic description satisfies the Markov hypothesis, or in other words, to quantify how much the conditional probabilities are history dependent. Given the transitions $\omega_j \to \omega_k$, if exists a finite time scale τ such that

$$T_{ijk}(\tau) = T_{jk}(\tau) + R_{ijk}(\tau) \text{ for any previous state } \omega_i \tag{2.105}$$

then, at that time scale all the probabilities are Markov with an error $R_{ijk}(\tau)$. It is clear that from the asymptotic properties of $T_{ijk}(\tau)$ and $T_{ij}(\tau)$ one has

$$R_{ijk}(0) = R_{ijk}(\infty) = 0 \tag{2.106}$$

which means that at the time zero and at the equilibrium[3] time scale the system is trivially Markov. Only for a finite time scale a mesoscopic description can be, until a certain extent, Markovian. On the other hand, the function $R_{ijk}(\tau)$ can tell how much non Markov are the transitions $\omega_i \to \omega_j \to \omega_k$. To check the condition 2.105 globally, we write the equation 2.105 in terms of the fluxes (using the relation 2.101), namely

$$P^e_{ijk}(\tau) = P^e_i T_{ij}(\tau) T_{jk}(\tau) + P^e_{ij} R_{ijk}(\tau) \tag{2.107}$$

Summing up over all the observed triple transitions we obtain the total non-Markov flux at the time scale τ as

$$\begin{aligned} F(\tau) &= \sum_{\omega_i \to \omega_j \to \omega_k} P^e_{ij} R_{ijk}(\tau) \\ &= 1 - \sum_{\omega_i \to \omega_j \to \omega_k} P^e_i T_{ij}(\tau) T_{jk}(\tau) \end{aligned} \tag{2.108}$$

This function depends on the number of available mesostates N and on the lag time τ, since the theoretical number of triple transitions approach N^3 for increasing lag times while its actual number is $M - 2 \ll N^3$ (M is the trajectory length). In particular the absolute number of transitions naturally grows when the lag time increase (at lag times comparable with equilibrium time scales all the transitions are theoretically possible). A too large value of the flux $F(\tau)$ generates an evolution of the master equation in which not observed transitions result in an artificial description of the mechanisms underlying the process studied. That because errors in the transition matrix may propagate on the master equation on long times extrapolations. In figure 2.22 we show the values of the non Markov flux F as a function of the number of steps (the lag time τ) at which the transition matrix is estimated. We computed all the transition matrices for lag times running from a single step to 10^4 steps (200 ns) for the mesoscopic descriptions SNC, SSS, SRA and RMSD[1.5] respectively. All the fluxes monotonically grow with the number of steps for the reasons previously accounted. The SNC mesostates are the worst in terms of markovianity, already at one step lag time about the 60 % of the total flux is non Markov. The situation improves for SSS (about 30 %) and is under the 20 % for the SRA and RMSD[1.5] descriptions. Last two methods appear essentially the same when examined for their Markov property. However a 20 % error is too large to be safely used. That is essentially due to the large number of mesostates, and to the fact that not all the mesostates share similar Markov property: there may be mesostates totally Markov or other mesostates, typically low populated ones, that are totally history dependent due to their poor statistics (that has been also noticed in [Park and Pande, 2006]). Thus, it is crucial to reduce the number of states to be used within this approach, namely to find a method able to redefine the space of suitable mesostates. That is the aim of an algorithm that we have developed and called "causal grouping" of the mesostates.

2.7.3 Causal grouping of the mesostates

We have seen that mesoscopic description of the conformational space, although they can classify large ensemble of microstates sampled in a MD simulation, their large number is not compatible with a

[3] Equilibrium here is meant for thermodynamic limit rather than convergent simulations. The difference is crucial since the latter deals with the finite size effects of the finite length MD trajectories.

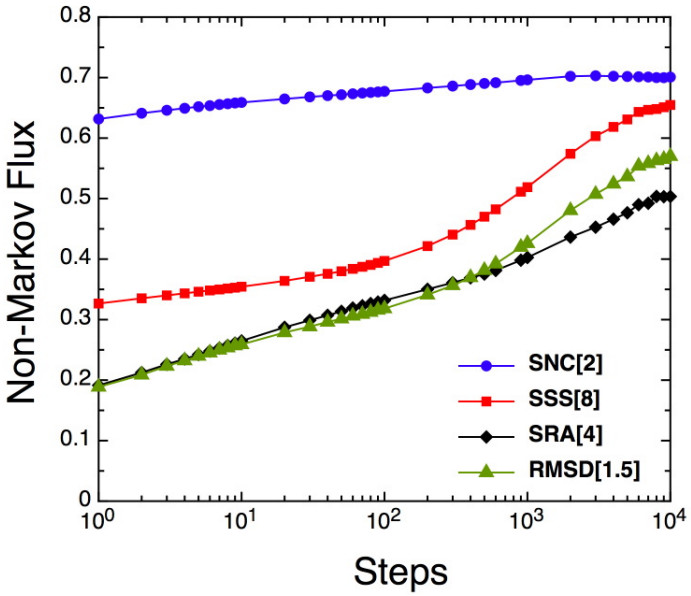

Figure 2.22: The amount of non Markov fluxes as a function of the number of steps at which the transition matrix T is estimated for 4 kind of mesoscopic descriptions of the configurational space of the GSGS.

Markovian description of the kinetics. Moreover in section 2.4.4 it was shown that only a minority percentage of the total number of mesostates are statistical relevant and stable, while their overwhelming majority can be considered as thermal fluctuations. In a Markovian description of a process the dynamics must proceed only between stable or marginally stable states, namely states whose interior the system has enough time to loose the memory of its past. Unstable states conversely proceeds along a time direction that is entirely determined from their past dynamical history that begins from the last visited stable state and ends up to the next stable state [Scoppola, 1993, Olivieri and Scoppola, 1996, Bonetto and Gallavotti, 1997]. To put it differently, Markovian mesostates satisfy to a diffusive dynamics while non-Markovian mesostates are ballistic. With these premises in mind we developed a mesostate reallocation algorithm based on the causality of the groups of mesostates that are statistically insignificant

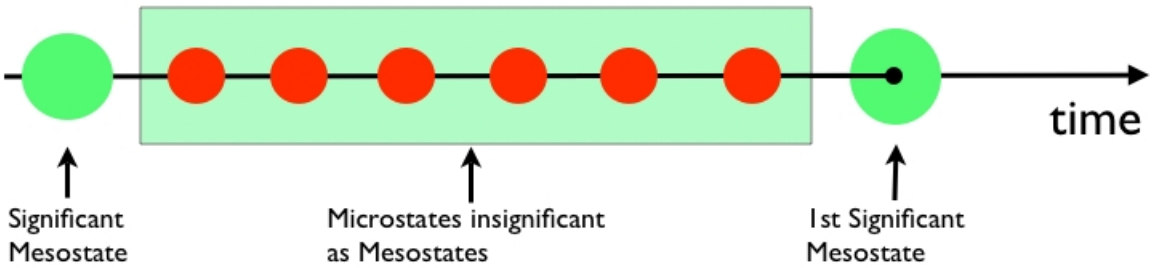

Figure 2.23: The idea behind the causal grouping algorithm.

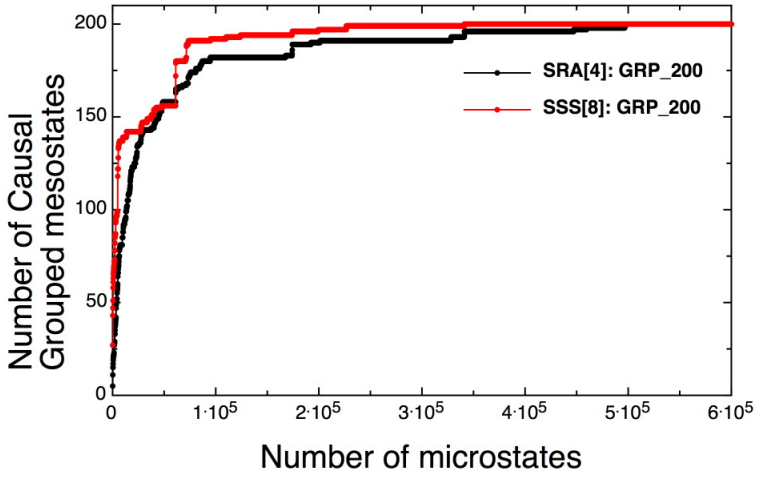

Figure 2.24: The convergence of the sampled number of causal grouped mesostates as a function of the length of the input simulation with a cutoff of 200 ($6 \cdot 10^5$ microstates corresponds to 12 μs of simulation time).

according to section 2.4.4. The word "causality" is properly chosen as the mesostates that are considered statistically insignificant are treated as microstates, and reassigned to the significant mesostates according to the "futures" they produce. Essentially the algorithm produces a new "filtered" time series of mesostates from the time series of "unfiltered" mesostates. Let us make an example, suppose we have found that in our mesoscopic description only about the first 300 mesostates are statistically significant while all the rest are declassed to microstates due to their low populations. Now (see figure 2.23 for a visual explanation), let's look to the time series of unfiltered mesostates at the time t, suppose the actual mesostate significant, then that is kept as such in the new filtered time series; let's suppose now that between the time window $t+1$ and $t+\Delta t$ there are insignificant mesostates and that at time $t+\Delta t+1$ a significant mesostate is encountered. Thus, all the mesostates in the time range $t+1$ and $t+\Delta t$ are considered as microstates and reassigned to their next significant future to the new filtered time series, namely the mesostate at time $t+\Delta t+1$. It is clear that such an algorithm corresponds to a sort of clustering procedure whose results depend on a cutoff, that is the number of significant mesostates. Therefore, the resulting filtered time series posses a number of mesostates corresponding to those that are significant with modified populations as a consequence of the reallocation procedure. Let us focus on two methods of mesoscopic description, SRA[4] and SSS[8]. From table 2.6 we established that the number of statistically significant mesostates for SRA[4] and SSS[8] are in the order of magnitude of 262 and 149, respectively. For comparison of the two descriptions we have chosen a cutoff of 200 between the two order of magnitudes. In figure 2.24 the performance of the algorithm, meaning the convergence in the number of sampled causal grouped mesostates is shown. For SRA[4] convergence is smooth and reach the 90 % of the total number of states already after $1/6$ of the total simulation length (i.e. about 2 μs). For SSS[8] convergence is more bursting and irregular though it is faster than SRA[4]. That could be due to the fact that SSS[8] is intrinsically finer grained than SRA[4] so that certain pathways might be better defined in SSS (not mixed up with some others) than SRA. Causal grouped mesostates essentially

depend on the type of mesostates the previous coarse graining generated. As we learned so far there are typically two kind of stable mesostates, those stabilized by the energy versus those stabilized by the conformational entropy. The effect of a conformational coarse graining on the free energy of a mesostate ω_i produces the relation $\Delta G_i = \Delta E_i - T\Delta S_i^{\mathrm{b}}$ where ΔE_i is the effective energy difference of a mesostate with respect to the ensemble mean energy, and ΔS_i^{b} is the internal entropy difference of a mesostate computed from the energy fluctuations of mesostate (see section 2.4.1), that is the entropy due mainly to the vibrational modes of the side chains and in some extent to the backbone. Consequently, the free energy per causal grouped mesostate has to take into account the conformational entropy due to fact that different configurational mesostates may be included into the same causal mesostate. That can be estimated from the Shannon entropy of a set of strings as we learned in section 2.5.1, namely the free energy per causal mesostate turns out to be

$$\Delta G_i^{\mathrm{causal}} = \Delta E_i^{causal} - T\Delta S_i^{\mathrm{causal}} - T\Delta h_i^{causal} \tag{2.109}$$

where Δh_i^{causal} is the conformational entropy loss computed on the ensemble of strings belonging to the causal mesostate (see equation 2.62).

We have introduced the causal grouping procedure to find a set of mesostates satisfying the Markov property. Focusing on the mesoscopic descriptions SRA[4] and SSS[8] we calculated the non-Markov flux for three cutoffs adopted for the causal grouping: respectively 200, 300 and 1000 mesostates. The cutoffs are the total number of causal grouped mesostates that result in the new filtered time series. In figure 2.25 the results of this analysis are shown. At the smallest time step (one step corresponding to 0.02 ns) the amount of the non-Markov flux is drastically reduced with respect to the non causal grouped mesostates. For a cutoff of 200 mesostates and $\tau = 0.02$ ns the SRA description produces less than 1% non-Markov flux of the total produced flux. The amount weakly grows until a convergence value of about 40 % is reached around at 200 ns time scale. The initial values grows with the increasing of the cutoff used for the causal grouping. In comparison with SRA, the SSS causal grouped description posses a higher initial value of non-Markov flux suggesting that descriptions based on torsional angles are more suitable for a Markovian treatment. However SSS initial value of non-Markov flux corresponding to a cutoff of 200 is less than the 5% of the total flux which is still a very low value compared to the non causal grouped description. Low cutoff values curves of figure 2.25 reach a convergence value which correspond to the equilibrium of the time series.

We previously pointed out that at the equilibrium the system is trivially Markov while here we obtained that at the equilibrium the non-Markov flux is maximized. Is this a paradox? Actually the thermodynamic equilibrium is intrinsically different from its equivalent in a simulation. At the thermal equilibrium all the states are theoretically accessible and all the transitions are possible for diffusive flight, namely the system is completely memoryless. In a simulation although one can find a way to define states that converge in their number (for instance the causal grouping), sampling the transitions is a more delicate issue since that is related to the finite size character of the simulations, and that is especially true for long time scales. In other words, what we estimate as non-Markov flux corresponds to the transition flux, predicted from the modeled Markov process, that has never been observed in the original time series. While at low time scales this flux in fact corresponds to a non-Markovian flux, for long time scales this evaluates the finite size effects of the the trajectory, namely the equilibrium flux for transitions that have never been observed in the trajectory. Thus, using low time scales to estimate a

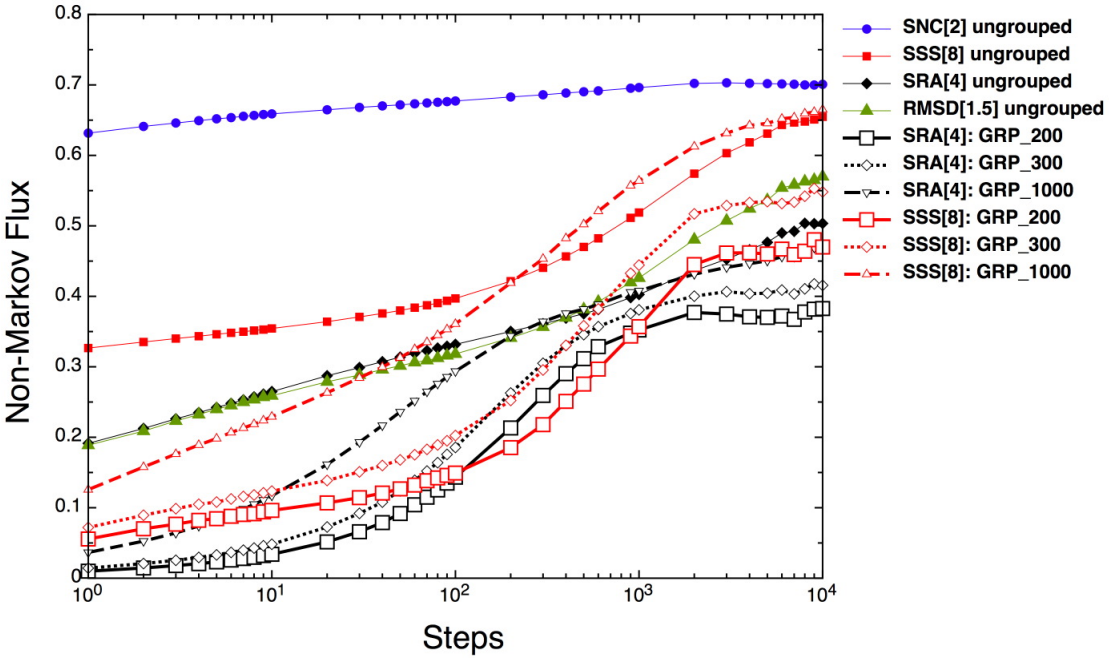

Figure 2.25: The non-Markov fluxes as a function of the number of steps at which the transition matrix T is estimated for the mesoscopic descriptions SRA and SSS respectively for a causal grouping at 200, 300 and 1000 mesostates in comparison with the non-Markov fluxes of the ungrouped descriptions SNC[2], SSS[8], SRA[4] and RMSD[1.5].

transition matrix on a proper set of mesostates allows to avoid the finite size effects of the time series or, putting it differently, the Markov property can be tested only on short time scales where the full statistics of the simulation can be employed and the finite size effects can be neglected. The algorithm presented has certainly two main strength points: its simplicity (it can effectively coded in awk), and also that the generated number of causal mesostates is an input variable of the algorithm. Although at low values of the cutoff the Markov property is favoured, too low cutoffs may produce causal mesostates that might loose any relation with the structural motifs of the polypeptide chain.

2.7.4 A Markov chain on the causal grouped mesostates

Once that a suitable set of mesostates are chosen, those defined by the causal grouping procedure, the master equation 2.92 can be readily applied as a tool of investigation. In particular a transition matrix T_{ij} can be estimated from the time series produced by the causal grouping procedure. Because we are dealing with a discrete time step, it is more convenient, instead of using a continuous master equation, to define a Markov chain based on the transition matrix T_{ij}. Given N mesostates and the population vector $\mathbf{P}(t) = (P_1(t), \cdots, P_N(t))$ at the time $t = m\tau$ a Markov chain is completely defined by the relation

$$\mathbf{P}(t) = T(\tau)^m \mathbf{P}(0) \tag{2.110}$$

where $\mathbf{P}(0) = (P_1(0), \cdots, P_N(0))$ is the initial condition. The properties of a Markov chain derive directly

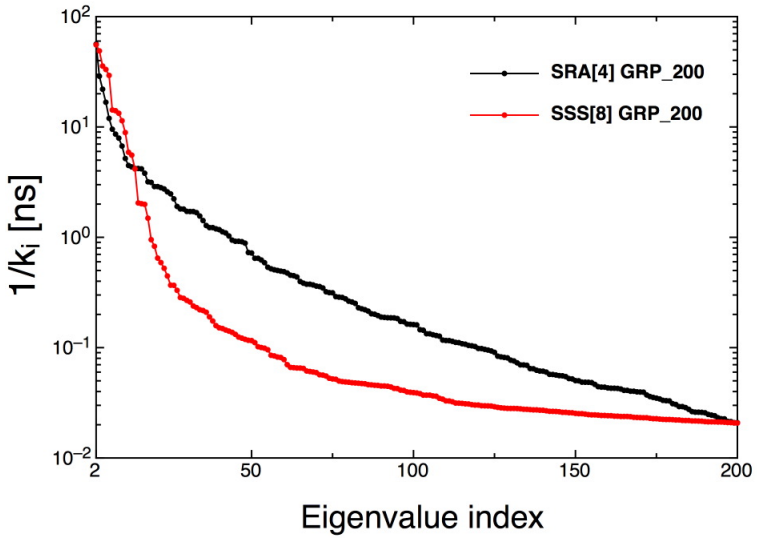

Figure 2.26: The relaxation time scales of the Markov chains computed from the inverse of the rate spectrum of the matrix $K = \mathbf{1} - T$ estimated on the causal grouped time series SRA[4] and SSS[8].

from the properties of its transition rate matrix T_{ij}. In particular the spectral analysis of T_{ij} determines the whole kinetic behaviour of the Markov chain. Let's consider the eigenvalue equation for T_{ij} so that

$$T\bar{\phi}_\lambda = \lambda\bar{\phi}_\lambda \tag{2.111}$$

with $\bar{\phi}_\lambda$ the eigenvector corresponding to the eigenvalue λ. To the steady state distribution $\mathbf{P}^e = \bar{\phi}_1$ corresponds the eigenvalue $\lambda = 1$, all the other eigenvalues are $\lambda < 1$. Any initial condition $\mathbf{P}(0)$ can be expressed as a linear combination of eigenvectors of the transition rate matrix so that

$$\mathbf{P}(0) = \mathbf{P}^e + \sum_{\lambda\neq 1} c_\lambda \bar{\phi}_\lambda \tag{2.112}$$

for some set of coefficients c_λ which must be chosen in order to keep the normalization of $\mathbf{P}(t)$ [van Kampen, 1981]. Thus, the equation 2.110 with the help of 2.111 reads

$$\begin{aligned}
\mathbf{P}(t) &= \mathbf{P}^e + \sum_{\lambda\neq 1} c_\lambda \lambda^t \bar{\phi}_\lambda \\
&= \mathbf{P}^e + \sum_{i=2}^{N} c_i (1-k_i)^t \bar{\phi}_i \\
&\simeq \mathbf{P}^e + \sum_{i=2}^{N} c_i e^{-k_i t} \bar{\phi}_i \ \text{ for } m \to \infty
\end{aligned} \tag{2.113}$$

where we have defined the rates $k_i = 1 - \lambda_i$ which are the eigenvalues of the matrix $K = \mathbf{1} - T$ with $\mathbf{1}$ the identity $N \times N$ matrix (note that $k_1 = 0$). The equation 2.113, for asymptotical times, correctly leads to the steady state distribution $\mathbf{P}^e$. The rates k_i represent all the relaxation modes of the Markov chain, where the modes are linear combinations of the chain states, that are the mesoscopic configurational states of the polypeptide. Therefore, the modes represent distributions of different configurational

states that lead to an overall relaxation rate. We have constructed a Markov chain with the transition matrices estimated from the causal grouped time series of the GSGS simulations using the mesoscopic time series based on SRA[4] and SSS[8] and a cutoff fixed at 200 states. In figure 2.26 the inverse of the rate spectrum is shown for both the Markov chain constructed from the causal grouped mesostates SRA[4] and SSS[8]. They represent all the time scales of all the relaxation modes to the chain equilibrium: from ∞ to a time step $\tau = 0.02$ ns. The second largest time scale (that corresponding to k_2) is the slowest mode which mainly corresponds to folding, interestingly the time scales for that turns out to be about 56 ns for SRA[4] and 55 ns for SSS[8] which is about the half of that found directly on the mesoscopic time series though the order of magnitudes are definitely comparable. There is a qualitative difference between the two spectra, for SRA the time scales decrease smoothly suggesting that the folded and unfolded phases are not clearly kinetically separated while for SSS the time scales decrease abruptly within few modes suggesting that an unfolded phase might be more clearly kinetically separated from the folded. Analyzing the normalized eigenvector corresponding to the rate k_2 gives account of the composition of the ensemble of causal states that contribute to the overall slowest relaxation mode $1/k_2$. An element of the normalized eigenvector represents the relative population of a macrostate contributing to that relaxation mode. The dominant contribution to the slowest relaxation mode is due to the folding reaction. Since the folding reaction starts from an unfolded phase, the eigenvector corresponding to the rate k_2 can be viewed as a structural model of the unfolded state for the GSGS peptide. In figure 2.27 we show the normalized eigenvectors corresponding to k_2 for the Markov chains based on the causal grouped states SRA[4] and SSS[8]. The intensity peaks on the y axes provide the relative population of the state with id i within the phase possessing the relaxation rate k_2. In other words these populations can be interpreted as the actual composition of the unfolded state of the GSGS. For some representative peaks we show their structural content in terms of an ensemble representation of the structures. All the structures are characterized by mainly four structural motifs: a curly like motif in which the second GSGS harpin is formed, a helix like motif (full helix, N-term helix turn extended, C-term helix turn extended), central helix and N a C terms interacting, a random coil motif characterized by semi-compact low energy structures. Essentially both the descriptions SRA[4] and SSS[8] give similar structural results. By looking to the mean energies and mean configurational entropies per causal mesostate we noticed that those mesostates contributing the most to the eigenvector posses generally low energy, comparable with that of the folded state. Other mesostates with lower intensity peaks seem generally to be stabilized by high conformational entropy. That is reasonable since one expects that those states acting as kinetic traps contribute the most in making the overall folding rate slower.

2.7.5 Network representation of the transition matrix

The transition matrix $T_{ij}(\tau)$ describes a mesoscopic dynamics between N causal grouped mesostates which approximatively satisfy to the Markov property; the term "mesoscopic" is correct because the time scale $\tau = 0.02$ ns is not a really microscopic (that is by instead the time step of the MD simulation which is usually 2 fs) neither macroscopic. Visualizing the transition matrix can be useful to get insight on what are the dominant system pathways or to get informations on the organization of the protein conformational landscape. In the group of Caflisch many efforts have been spent in the last years in the direction of a network representation of the folding process in light to provide a so called

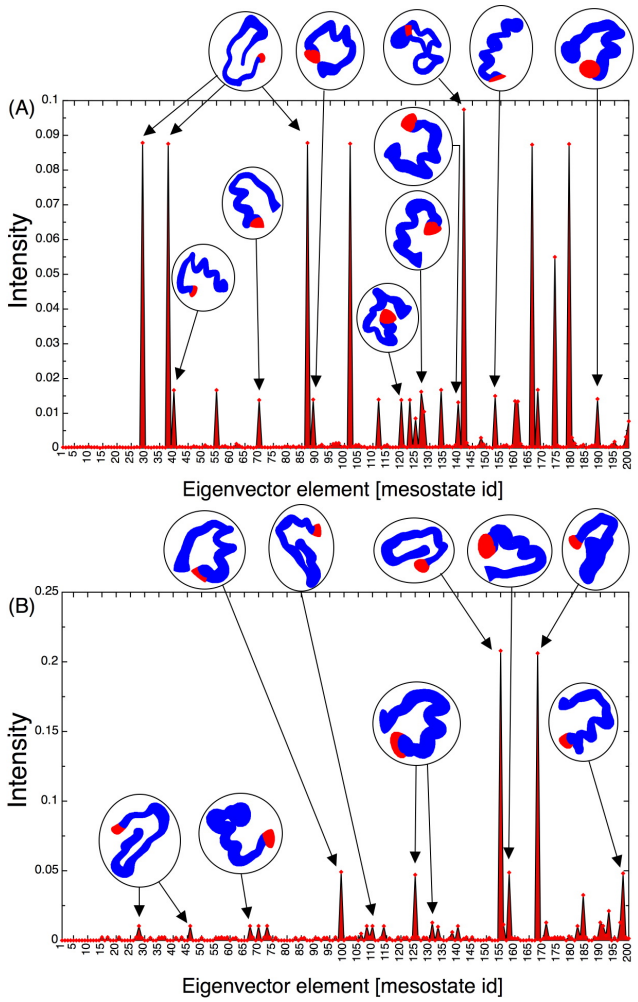

Figure 2.27: The normalized eigenvector corresponding to the slowest non null rate k_2 of the Markov chains constructed on the causal mesostates SRA[4] (A) and SSS[8] (B). Some intensity peaks are shown together with the ensemble of structures they correspond to. Pictures have been made with the program molmol.

"unprojected" view of the system free energy landscape [Rao and Caflisch, 2004, Caflisch, 2006, Gfeller et al., 2007]. Complex network representation is a powerful tool to elucidate correlations and interdependencies among the species of a complex dynamical system. They are widely used from sociological network studies [Watts and Strogatz, 1998] to trade economy [Shapiro and Varian, 1999] passing through the study of the internet evolution and structure [Yook et al., 2002] and ending up to the metabolic network in cell biology [Jeong et al., 2000]. In our context we use the network representation to visualize the mesoscopic dynamics associated to the Markov chain with the aim of clarifying how are the dominant folding pathway and whether or not there is intercross dynamics between the free energy basins composing the unfolded state. Our reference descriptions are those of the causal grouped mesostates with 200 and 1000 mesostates for SRA[4] and SSS[8] depending on representation convenience. All the transition matrix graphs are visualized using the open source program Tulip [Auber, 2003]. Many visualizing

algorithms exist to represent large graphs, here we have chosen that which is more close to what one expects a free energy landscape can look like. The algorithm is implemented in Tulip and it is called GEM (Graph *em*bedder) which is an extension on the spring-embedder approach (see [Frick et al., 1994, Bruß and Frick, 1995] for details). Spring-embedder algorithms use a physical model based on forces that are exerted on the vertices in order to improve their positions according to several aesthetics. Once the vertices are placed, the edges are drawn as straight lines between the vertices. The model states that far vertices in terms of connectivity repel each other, while adjacent vertices are attracted to each other. These simple rules define a dynamic system that can be driven into a local energy minimum. The easiest strategy to find out such a minima is to use a gradient descent method, according to which only downhill moves are allowed, until no further improvements are possible. Other strategies of achieving convergence are the use of simple cooling schedules that restrict the allowed moves over time, or to apply simulated annealing. In particular the GEM Frick algorithm combines the spring-embedder approach with the ideas from simulated annealing by assigning each vertex a local temperature. This algorithm turns out to be very effective in terms of running speed as well as the robustness of the displacement found (meaning that if one runs twice the algorithm on the same graph the same displacement is found).

In figures 2.28 and 2.29 the transition matrix graphs for the causal grouped mesostates SRA[4] and SSS[8] with 1000 vertices are shown respectively. The number of edges visualized are 15608 for SRA[4] and 33760 for SSS[8], no cutoff on the transition probabilities has been used to facilitate the visualization so the graphs represent entirely the transition matrices as estimated from the causal grouped time series. For the SRA[4] network there is a clear separation between the helical phase and the beta phase. Helical phase appears to be formed by four helical structural motifs that we called H1, H2, H3, H4 that sum up to about 7 % of the total statistical weight. The β phase has in its center the full folded mesostate that is surrounded by other low populated mesostates which sum up approximatively about more than 60 % of the total weight. The curl like basins C1, C2, C3 represent kinetic traps that overall on this description sum up to about 5 % of the global weight. On the periphery of the folded state there are two interesting basins E1 and E2 both characterized by the beta turns already formed but very fluctuating in the rest of the topology. These basins weights about 3 % and 2 % respectively. They are virtually included in the folded basing suggesting a possible their role as gateways or as on pathway intermediates to the folded state from both the helical phase and the kinetic traps. Thus the network of the causal grouped mesostates SRA[4] with 1000 vertices suggests there are three main classes of basins in the free energy landscape: a helical basin, a beta trap basin and broad folded basin.
The network corresponding to SSS[8] of figure 2.29 is different from the previous. It appears much more detailed or even redundant compared with that of SRA[4]: many structural motifs are found in different basins, for example the curl like structures called C1, C2, C3, C4, C5, and also the triple stranded motif of the folded state is found in 4 basins, including the main folded one. The helical basin appears again well separated from the rest suggesting that this separation is a true phase difference not really something depending on the adopted description. The weight of the helical basin is again about 7 % while that of the β trap is about 6 %; the weight balance between these two basins with respect to those in SRA[4] is confirmed for the helix basin and overweighed for the curl-like basins. The folded basin is organized in sub-basins, the main one weights about 21 % while the other three 3s1, 3s2 and 3s3 have a total weight of about 15%. Clearly the fact that the strings based on secondary structure are base on an alphabet of

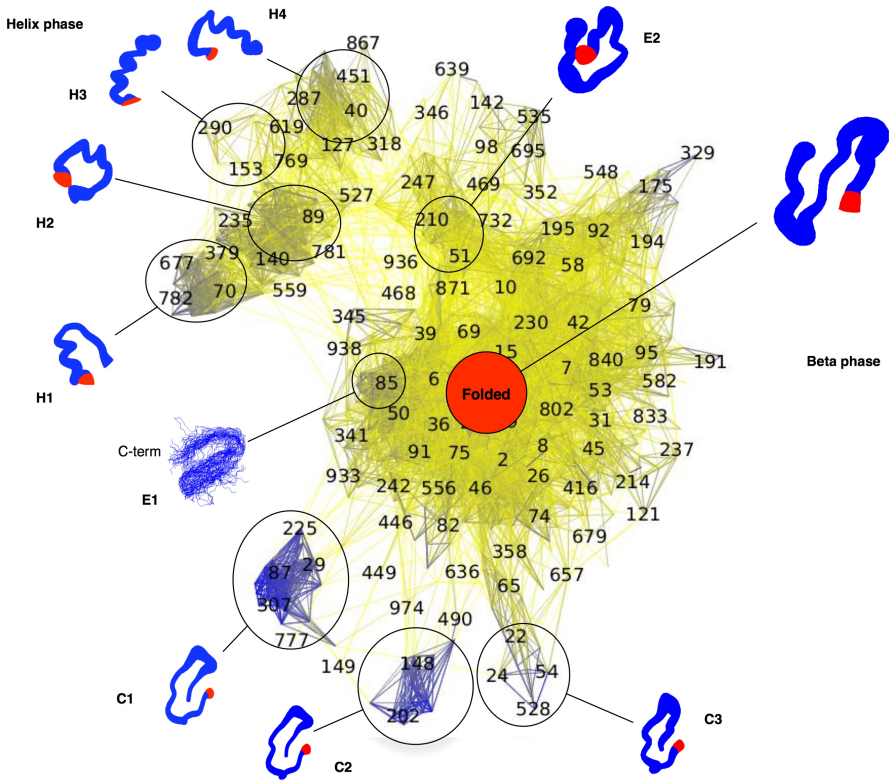

Figure 2.28: The network corresponding to the transition matrix estimated from 1000 causal grouped mesostates based on the SRA[4]. The total number of vertices are 1000 and the edges are 15608 without any cutoff on the probabilities. The graph has been realized with Tulip and the edges have been coloured according to the value they assume on the transition matrix. It appears clear that two main phases characterize the transition matrix: a helix phase clearly separated from a beta phase at whose center is posed the folded state.

8 symbols, the double than SRA at equal string length, makes the free energy landscape more detailed and thus, possibly redundant. Moreover the redundancy of the secondary structure description (also reflected in the higher Shannon entropy) might play an important role in the Markov property introducing memory effects.

The transition matrix corresponding to 200 causal mesostates are represented in figure 2.30 for both the descriptions SRA[4] and SSS[8]. The differences are evident. While the in the SRA graph the folded state appear concentrated in a rather unique basin and a helix basin is still separated from other structural motifs, for the SSS graph only β like basins are well defined. In particular along with the main folded state other three basins characterized by the triple stranded (we call them pre-folded) result to be in equilibrium with the main folded basin. There are two interesting mesostates in between the basins pre-folded1 and main folded, characterized by a strongly fluctuating N-term harpin: these states seem to play the role of transition states between sub-basins of the folded basin. No proper helix like basins appear, all the helices are located in isolated vertices in between the different beta basins, suggesting that in this description in the helical unfolded state the system diffuses without trapping itself. Thus

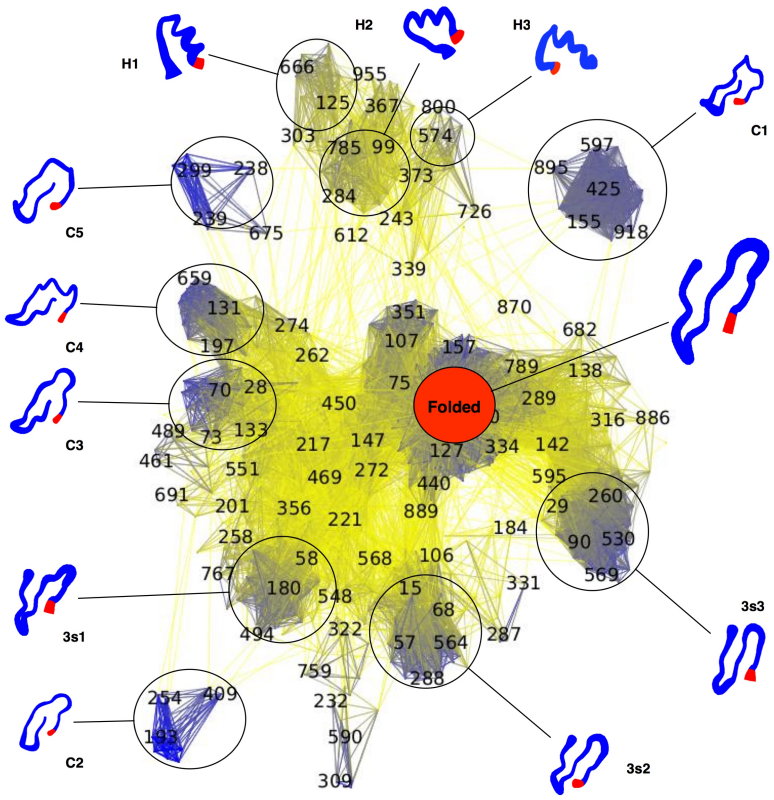

Figure 2.29: The network corresponding to the transition matrix estimated from 1000 causal grouped mesostates based on the SSS[4]. The total number of vertices are 1000 and the edges are 33760 without any cutoff on the probabilities. The graph has been realized with Tulip and the edges have been coloured according to the value they assume on the transition matrix.

, reminding that the graph pictures are the result of a visualizing algorithm, we can qualitatively conclude that at this stage of coarseness while a SRA description seems to organize the vertices (causal mesostates) in such a way that free energy basins naturally arise (both enthalpy or entropy driven) in the SSS description only basins characterized by low enthalpy arise. On the other hand from the graph pictures we are not able to infer precisely what are the sequence of events the lead to the folded state from any starting point, although free energy conformational basins appear spontaneously. We can get a pictorial idea on how a free energy landscape could look like, but we should not forget their strong subjectivity. However, an interesting question arising from the observation of the graphs is whether or not inter-conversions between unfolded basins on a time scale less than the folding time. An answer to this question might shed light on the folding mechanism: is the unfolded state made by basins that can inter-convert between them before getting into the main route to folding or no inter-convertion is possible and thus exclusive folding pathways exist? To these and to other kind of questions the Markov treatment of folding kinetics can answer.

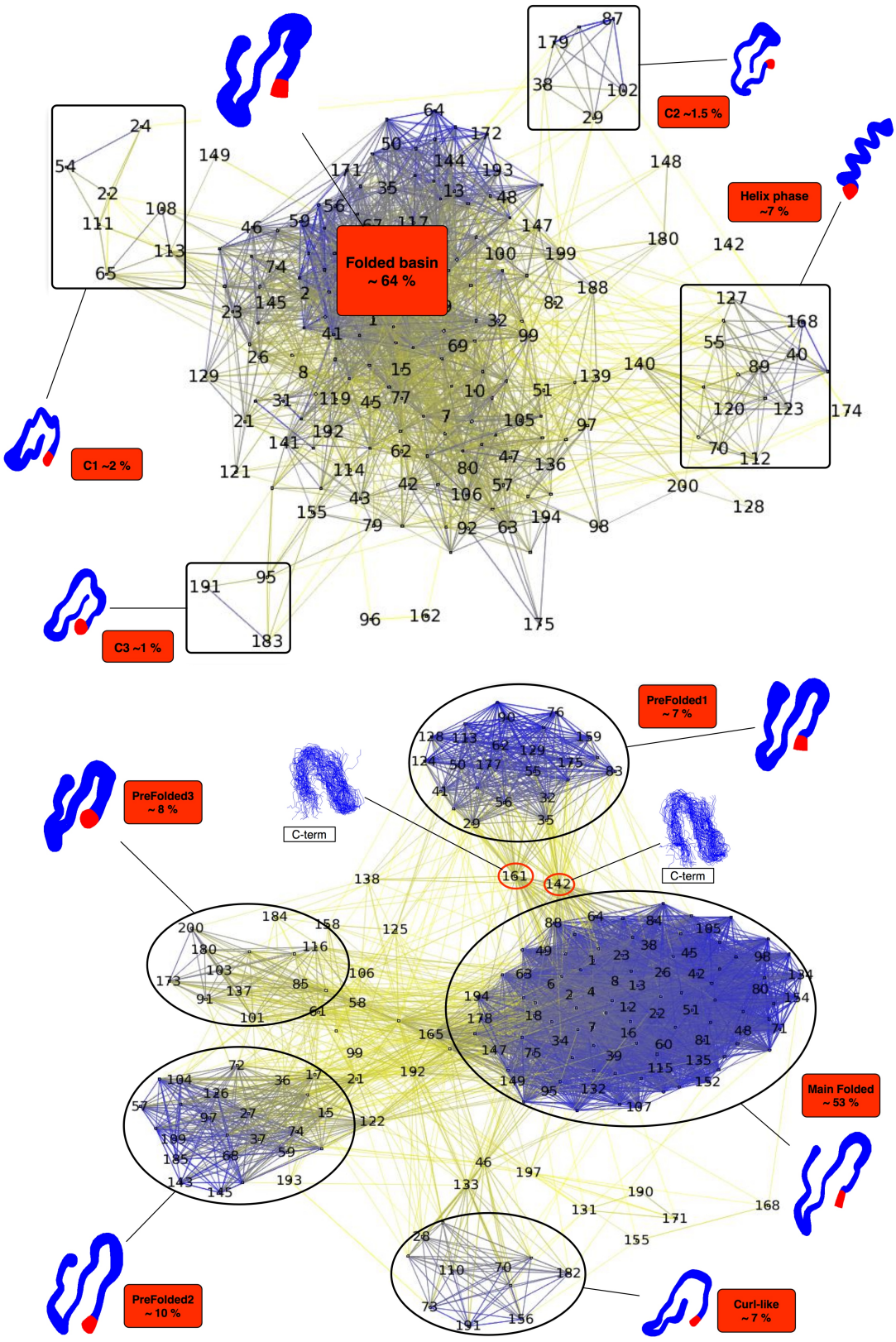

Figure 2.30: The transition matrix graphs corresponding to 200 causal mesostates for SRA[4] (top) and SSS[8] (bottom). Graphs have been realized with the program Tulip. Circles and boxes represent basins clearly distinguishable from other vertices. Indicative populations of the basins are reported. In the bottom network a helical basin is missing, helices are present as transient nodes (see the node 99 in the bottom network) as the rank of the first helix SSS[8] mesostate is much lower than that in SRA[4] (the 99th "`---HHHHHHHHHHS------`" for SSS[8] and 40th "`0100011111111111111`" for SRA[4])

2.7.6 MFPTs from a Markov chain

An effective way to investigate the folding pathways using the Markov approach is to calculate the inter-conversion MFPT from any causal mesostate to another. Thus given N causal mesostates with their mesoscopic transition matrix T_{ij} the idea is to calculate the equilibrium times that in average are necessary to reach any state starting from any another. That leads to the construction of a matrix M_{ij} of the mean first passage times. To calculate the MFPT matrix M_{ij} from the mesoscopic transition matrix T_{ij} a classical treatment for finite and ergodic Markov chains[4] can easily be adopted following [Snell, 1959, Kemeny and Snell, 1976]. The procedure is based on elementary linear algebra and requires to define some matrices at first and it is equivalent [Kemeny and Snell, 1976] to solve the following system of linear equations

$$\begin{aligned} M_{ij} &= \tau + \sum_{k \neq j} T_{ik}(\tau) M_{kj} \\ M_{ii} &= \sum_{k} T_{ik}(\tau)(M_{ki} + \tau) \end{aligned} \tag{2.114}$$

The transition matrix T defines an equilibrium matrix W_{ij} through

$$\lim_{n \to \infty} T^n(\tau) = W \tag{2.115}$$

in which each row is equal to the equilibrium populations of the N causal mesostates. The convergence was reached already with an exponent $n = 20000$ which corresponds to a convergence time of 400 ns. Secondly one has to define the *fundamental matrix* Z for ergodic Markov chains which is given as

$$Z = (\mathbf{1} - T + W)^{-1} \tag{2.116}$$

where $\mathbf{1}$ is the identity matrix. Finally the matrix of the MFPT is obtained by the formula

$$M = \tau(\mathbf{1} - Z + E\, Z_{dg})D \tag{2.117}$$

where D is a diagonal matrix with entries $1/W_{ii}$, E is a matrix with all 1's and Z_{dg} is the diagonal matrix built taking the diagonal of Z. The factor τ converts the number of steps into a value in ns.

We calculated the matrix M_{ij} for the Markov chain corresponding to the causal mesostates SRA[4] with 200 vertices. The mesostate with id 1 corresponds to the folded mesostate, therefore the first row of the matrix $M_{i\to 1}$ provides the MFPTs from any starting mesostate to the folded state. Yet, the main diagonal of the matrix $M_{i\to i}$ gives the so called mean recurrence times of the mesostate, namely the mean time necessary to a state to return back to itself. The recurrence times are a measure of kinetic stability of the mesostates since these times are proportional to the inverse of the exponential of the internal barrier that a mesostate has to jump to get outside. In line with the choice in section 2.6.2, we consider the MFPT to the folded state as a sort of reaction coordinate. Thus to facilitate the analysis of the M_{ij} matrix we reorder its indexes in such a way that the low indexes (from 1) are those mesostates possessing low MFPT to the folded state, while large indexes have larger values of MFPT to the folded state. We call the reordered matrix M^* so that the first row satisfies the inequalities $M^*_{1\to 1} \leqslant M^*_{2\to 1} \leqslant \cdots \leqslant M^*_{200\to 1}$. In figure 2.31 the reciprocal of the recurrence times $M^*_{i\to i}$ (left y axes) as a function of the mesostate index

[4] An ergodic Markov chain is one such that it is possible to go (not necessarily in one step) from any state to any other state.

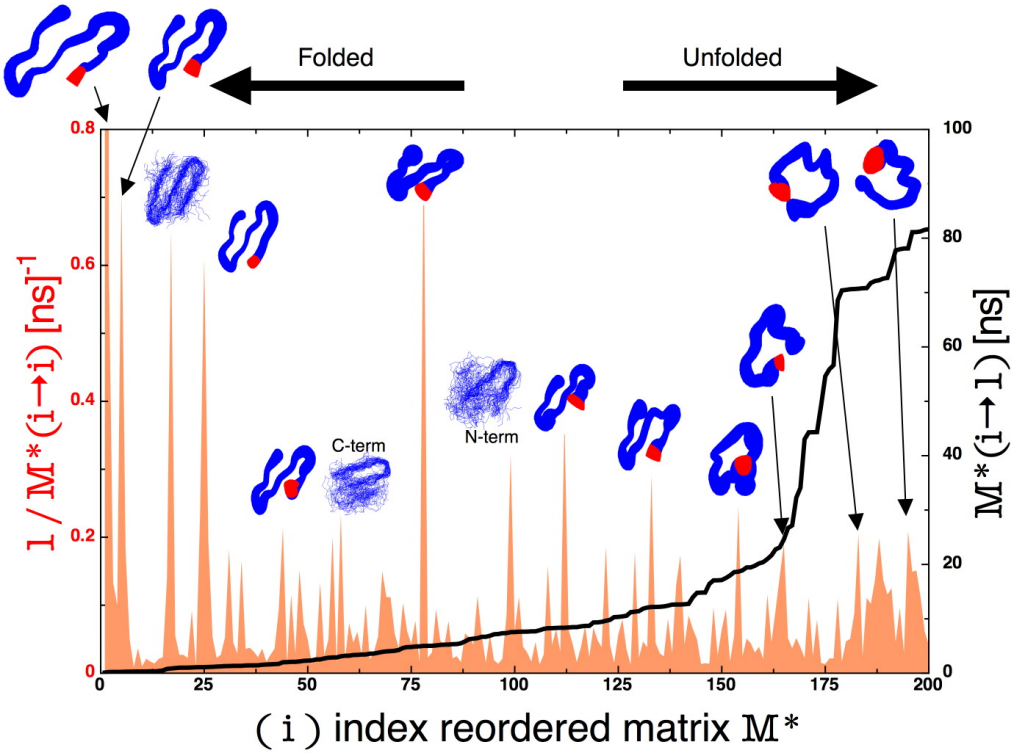

Figure 2.31: The inverse of the recurrence time $1/M^*_{i\to i}$ (red curve) referred to the left y axis; the black curve (right y axes) represents the MFPT to the folded state on the reordered MFPT matrix, namely its first row $M^*_{i\to 1}$. The x axes gives the reordered mesostate id from low MFPT to large MFPT to the folded state. Averaged structures are shown over representatives peaks of the recurrence time. Causal mesostates from SRA[4].

i is shown. The first row representing the MFPTs to the folded state $M^*_{i\to 1}$ is also shown in figure (right y axis). The peaks corresponding to the rates $1/M^*_{i\to i}$ represent kinetically stable causal mesostates. For some of these peaks, an ensemble view of the structures belonging to the mesostates is shown. The most unfolded mesostates in terms of their folding times (ranging from 50 to 85 ns) that are kinetically stable are helix-like mesostates. When the system is in the unfolded phase it tends to form stable helices in such a way the N and C terms can interact. By thermal fluctuations, helices can be disrupted while the β-turns can be easily formed. When the β-turns are accessed the system has two options: either forming a curl-like structure (external N-term) or going towards the folded state. The folded state is a huge basin which appears layered. There are many ways in which a triple stranded β-sheet can be arranged with an overall well defined topology and with a differentiated network of interactions. Many sub-basins are in equilibrium showing different amounts of fluctuations. The very bottom of the folded basin is enthalpy driven, others folded sub-basins are fluctuation driven so that, as a whole, the folded state can be thought as an entropy driven ensemble. In figure 2.32 we show the MFPT reordered matrix $M^*_{i\to j}$ (using causal mesostates from SRA[4]). Indexes i greater than 150 are referred to initial mesostates which can be already considered unfolded as one can see from figure 2.31. The horizontal bands (j index fixed and i index variable) represents ensemble of initial mesostates i relaxing to the mesostate j. All the

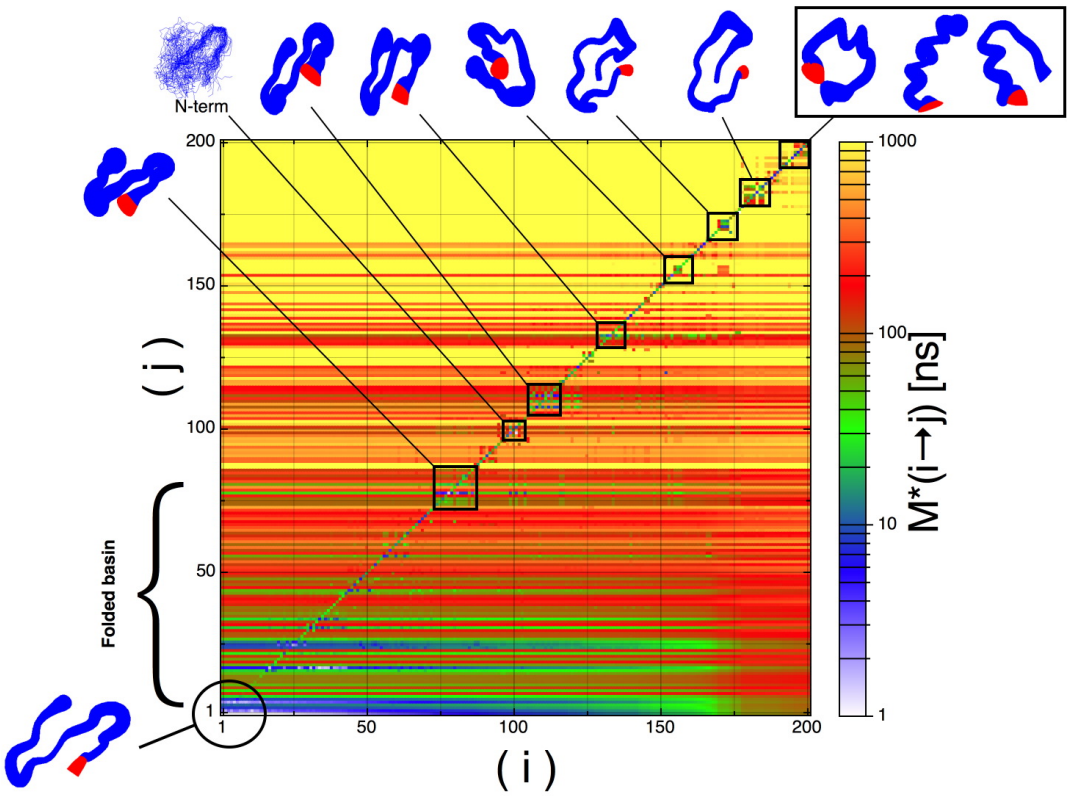

Figure 2.32: The reordered MFPT matrix $M^*_{i\to j}$ shows that the native basin works as a "hub" in the overall kinetics. An entry on the matrix gives the MFPT for the equilibrium transition $i \to j$ from causal grouped mesostates SRA[4]. Horizontal bands are equilibrium transitions from all the is to a specific j. Yellow points can be assumed to have an infinite barrier between initial and final state so that their inter-conversion rate is zero. Helix-like and curl-like do not exchange between them in a time scale that is comparable with their folding time. They directly relax to to folded state in about 90 ns and 60 ns respectively.

yellow entries are transitions that occur in a time scale $\gg 100$ ns: thus since this time is quite larger than the folding time an infinite barrier can be assumed between these states. The most important unfolded basins are the helix-like and the curl-like. According to the MFPT matrix we can establish that these two basins do not inter-convert between them in a time scale comparable with their folding time. From the dark bands on the bottom of figure 2.32 we can see that these basins directly relax towards the folded basin which begins from the index $\sim$80 in the matrix M^*. The folded basin is layered and can be acceded from different gateways which are characterized by strong fluctuations. Moreover the "kinetic radius" of the folded basin is about 20 ns, which is a time scale consistent with that found for the pre-folded phase by means of FPT calculations (see section 2.6.1). Thus the folding mechanism of the GSGS follows from the MFPT matrix: the helix phase is the real unfolded state which is corrugated and relax directly to the folded basin in about 90 ns; the curl-like phase can be seen as an off-pathway intermediate or a death route, when the system ends up there it needs to come back in order to find the folded basin. However, finding the folding basin from a curl-like topology is easier than from the helix phase because

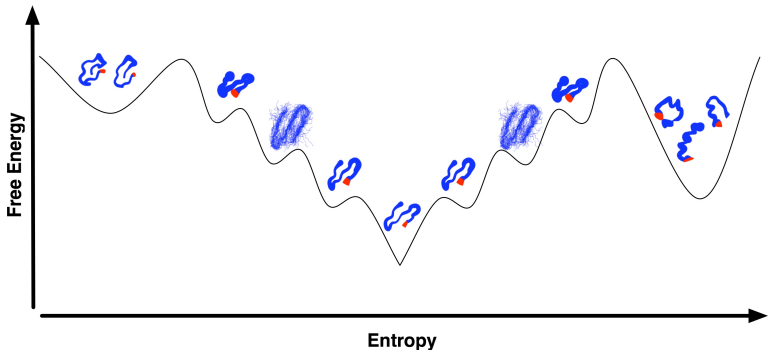

Figure 2.33: The free energy landscape emerging from the Markovian treatment of the GSGS kinetics. Low entropy states (curl-like states) act as kinetic traps while the helical-phase represents the unfolded state that is stabilized by high entropy. The folded basin is stabilized by both enthalpy and entropy. Folding can then initiate either from low enthalpy or high entropy states.

the former posses already a native like secondary structure and a hairpin already formed.

2.8 Conclusions

The *in silico* properties of the 20-residue GSGS peptide have been broadly analyzed, mostly by Caflisch and collaborators. This peptide constitutes and ideal model because the experimental behavior (reversible folding to a triple stranded β sheet) can be reproduced *in silico*, in a broad interval of temperatures, and with various physical, transferable force-fields. In particular, most simulations and those here analyzed have been performed with a particularly simple implicit solvation model [Ferrara et al., 2002], which is computationally very efficient, and works very well for the purpose, i.e., it reproduces the behavior expected from a two-state folder with a stable ordered phase. Despite its simplicity, this system revealed an underlying complexity and richness which also depends much on the scale at which it is observed. Recent results based on network analysis of the trajectory [Caflisch, 2006] confirmed the heterogeneity of the denatured ensemble which appear evident with the multiple method analysis presented here in this chapter. The unfolded ensemble contains not only ordered low enthalpy conformations, but also high enthalpy states which are stabilized by a high entropy. Moreover, the basins shaping the unfolded phase do not inter-convert between them on a time scale comparable with that of folding which suggests a multiple pathway paradigm for folding. Several methods have been adopted to describe the configurational space of the peptide, and different paradigms have been chosen for the thermodynamic and kinetic analysis. In particular, an attempt to elucidate the physical meaning of the coarse graining for polypeptide chains has been carried on. The consequences on the thermodynamics due to a choice of mesoscopic descriptors were analyzed in terms of the informational content of the mesostates. Mesoscopic descriptions in terms of symbolic states (strings) have revealed that the configurational space of a peptide can be analyzed with methods that resemble linguistics. Thanks to that kind of analysis one can establish that only an amazing minority amount of mesostates can only be con-

sidered statistically significant to represent the peptide configurational space while the overwhelming majority of mesostates can be considered fluctuations. By virtue of these analysis a hierarchical organization of the configurational space emerges. Notably, extremely simple models based on sigle residue state description give reasonably account on how the folded state is organized both thermodynamically and kinetically (the modified Zwanzig model) providing realistic folding times. Most importantly, a Markovian treatment of the configurational dynamics has been developed. The problem of Markovianity associated to a coarse-graining has been studied and partially solved by introducing a reallocation algorithm of the mesostates based on their causality. Markov treatment allowed to use a master equation to fully study folding kinetics. Thanks to this treatment a rigorous analysis of the configurational space of a polipeptide kinetics appears possible. Folding process is then treated as a stochastic reaction in which different configurational species are involved. From the mesoscopic trasition matrices trough an ergodic Markov chain we have studied equilibrium pathways and found that the GSGS folding process is determined by free energy landscape resembles that of figure 2.33, where a hypothetical reaction coordinate based on configurational entropy is used for qualitative considerations. The folded basin is layered by partially folded and fluctuating sub-basins so that it results stabilized by a both enthalpy and entropy. At the borders of the free energy profile two unfolded basins do not inter-convert and fold independently with different folding times. Notably, the left basin (curl-like) is characterized by low enthalpy and low entropy and plays the role of a kinetic trap while the right basin (helix-like) is mainly stabilized by high entropy. The estimated folding time from Markovian treatment is robust and comparable with that computed directly on the trajectories time series. In the next chapter 3 the methods so far presented and applied on the GSGS peptide will be used for the study of larger polypeptides based on simplified amino acid sequences.

3 Simulations studies on simplified protein sequences

3.1 Introduction

How many residues types are necessary to fold a protein? How minimally large should be an amino acid alphabet such that a protein sequence is foldable? These questions have been often posed in the context of protein design [Shakhnovich, 1998, Buchler and Goldstein, 1999b, Shakhnovich, 2006] with the aim to understand how natural proteins have evolved and how proteins can be engineered to perform novel functions. In this chapter we are interested to investigate this issue under a fully computational view point. The use of computer simulations to study folding mechanisms is limited to relatively small polypeptides. That is because the folding time of proteins generally scales both on their chain length [Gutin et al., 1996] and on the topology of the folded structures [Plaxco et al., 1998]. Thus if the folding times in the protein universe generally scales from the microseconds to the seconds (see e.g. the Protein Folding Database http://pfd.med.monash.edu.au), while the computational time to study a 20 residue peptide is about a month for 10 μs equilibrium simulations [Ferrara and Caflisch, 2000], it is clear that for small proteins (60 amino acids) the computational time would be prohibitive. Moreover one should take into account the fact that modern all atom force fields are not free from errors. Both underestimation and overestimation of the strength of specific non-bonded interactions may lead to the increasing of the free energy frustration of the modeled protein, with a consequent dramatical decrease of the folding rate. In other words, even if a force field were able to predict the correct folded state of a protein the time needed to computationally demonstrate it would be on the Levinthal time scale. This apparent pessimistic conclusion actually does not really depend on the technological evolution of the computers, it rather concerns the intrinsic nature of folding and what kind of models can be adopted to describe it. The questions we would like to address here are: how can a protein sequence be simplified so that one can computational increase its folding rate? What is the minimal number of amino acid letters which allows us to model a known structure and observing reversible folding *in silico*? If on one hand these questions lead us to make computationally tractable the study of folding for small proteins, on the other hand, such a study is directed towards the very nature of protein folding, the relation between proteins and the evolution of amino acids alphabets and eventually the correspondence between protein function and native structures and the relationship between folding and sequence information. On this respect, the experimental studies conducted by Davidson and coworkers on random libraries of sequences with only three amino acids constitute a remarkable starting basis [Davidson and Sauer, 1994, Davidson et al., 1995, Cordes et al., 1996]. In these studies a library of synthetic genes encoding 80- to 100-residue composed mainly of random combinations of glutamine Q, leucine L, and arginine R were expressed in

Escherichia coli. Among the proteins obtained some (about the 1% on a huge library) QLR proteins were well expressed and well characterized. These proteins, although totally artificial, have been shown to posses high helical content from CD measurements. Moreover, denaturation studies have also shown that in certain cases (tuning the sequence hydrophobicity) their folding/unfolding mechanism can be assumed cooperative. The interpretation of the results is made on the observation that the QLR residues used, combined in a proper way, can give sequences whose hydrophobicity is comparable to that of natural proteins. These studies led Davidson and collaborators to suggest that the key elements of protein design seem to be the proper placement of hydrophobic residues along the polypeptide chain and the ability of these residues to form a well packed core. According to them buried polar interactions, turn and capping motifs and secondary structural propensities also contribute, although to a lesser extent.
Other important works on this line have been carried out in the group of Baker. Notably while in Davidson works the analyzed sequences were totally artificial in [Riddle et al., 1997] a β-sheet protein, the SH3 domain, was simplified by using 5 letter amino acids: Isoleucine I, Lysine K, Glutamic Acid E, Alanine A and Glycine G. The study was conducted using a phage-display selection strategy to promote the biological protein activity. The use of the residues I, K and E was justified by the fact that globular proteins contains non-polar interiors and polar exteriors so any experimental simplifying framework should contain both polar and non-polar residues. Alanine and Glycine were the better conserved residues in the combinatorial libraries. Despite the dramatic change in sequence, the folding rates of the simplified versions of the SH3 protein were very close to that of the wild type. Moreover NMR analysis shown a well packed core which justify the high protein stability. Thus the selection procedure eliminated molten globular structures in favor of function. Finally, Baker and coworkers argue that simplified sequences constitute an opportunity to investigate the evolution of the rapid and cooperative folding of small proteins. Protein function needs that the native state of proteins be both stable and kinetically accessible. While the former is clearly under evolutionary pressure it is still unclear whether the latter is also an evolutionary factor. In their study and elsewhere [Plaxco et al., 1998, Watters and Baker, 2004] it is stressed that the number of letters required to obtain a foldable sequence could not be lowered to 3, they were unable to obtain foldable sequence containing only one polar and two non-polar amino-acids. In his "As simple as can be?" [Wolynes, 1997] Wolynes discusses Baker's results saying they fit very well with the energy landscape ideas for folding. He states that although others have succeeded in designing a four helix bundle using only a 3 letter alphabet [Regan and De Grado, 1988], that likely holds only for highly symmetric folds while higher complexity is needed to encode the features or more exotic folds (such as the SH3 β-barrel). In his view, too simplified sequences would not have an enough "stability gap" , the energy difference between the native state and the rest of the configurational space, to assure thermodynamic control. In fact according with the energy landscape models the ratio between the energy ruggedness and the stability gap establishes whether folding is either under kinetic control (large ratio) or thermodynamic control (low ratio) [Bryngelson and Wolynes, 1987, Onuchic et al., 1997]. In his conclusions he suggest that particularly symmetric structures could be encoded in a 3 letter amino acid alphabet, a fact that would possibly bring into question the role of the hydrophobic code in protein folding.

In the following we attempt to demonstrate that a three residue alphabet based on the secondary structure amino acid propensities is able to encode the topologies of β and α/β mini proteins. The residues we have chosen are Alanine for α-helices, Glycine and/or Serine for turn/loops and Threonine

for β-sheets. In particular we use AGS sequences to construct a series of four β-sheet polypeptides respectively having 20, 28, 36, 44 residues which are composed by poly-threonine stretches interrupted by respectively 2, 3, 4, 5 glycine-serine loops. Such sequences are shown to spontaneously fold in β-sheets by MD simulations in which the starting conformations are full extended. Their folded state is respectively given by a three-, four-, five-, six-strended β-sheet. Moreover, it will be shown that the configurational space is smoothly divided in two well distinct phases, a beta phase and a helix phase respectively, the latter representing mainly the unfolded state.
We also construct an AGT sequence in order to simplify the sequence of the B1 domain of protein G, a 56 residue α/β protein. We show through MD simulation that this sequence-protein spontaneously folds onto the protein G topology also from a completely extended starting conformations. The folded state resembles a molten globule such that the native secondary structure is steadily formed while the folded topology is diffuse. Strikingly, folding kinetics is fast and strongly non-cooperative so that multiple reversible folding events can be observed *in silico*.

The results suggest that the evolution of the amino acid alphabet and consequently of sequences might have been modeled upon the pressure of seeking the protein stability and function. On this respect, an interesting metaphor is that of the evolution of the natural languages for which the optimization of semantic content vs. the rigidity of syntactic structures is thought to have been the key of their evolution [Lieberman et al., 2007]. Protein sequences might have been similarly evolved from a set of local rules on how to cast together amino acids (syntactic-structural level), to highly complex sequences that are able to encode functions. The two levels, structural and functional, possibly could be achieved with different length of the amino acid code. Not much dissimilar arguments can be found in a recent work of Rose a coworkers [Fleming et al., 2006] in which using secondary structure constraints, the native folds of several proteins are reconstructed through a Monte Carlo annealing. Baker and collaborators have studied a fully computationally designed α/β protein (Top7) which shows a significant less cooperative folding than equivalent similar size natural proteins [Watters et al., 2007]. They argue that the cooperative folding of small size natural proteins is likely to be not a general property of polypeptide chains but instead a consequence of sequence evolution. Our results are along this line, low complexity amino acid alphabets on one hand may encode particularly symmetric topologies, on the other they lack of the necessary sequence redundancy and specificity which ensure protein stability and function.

3.2 Simplifying strategies

Many strategies have been introduced to develop a simplifying scheme of the amino acid alphabet based on the observation that the full sequence complexity is not required to encode the structural information of a protein so that amino acids can be simplified according to their physical-chemical properties (see [Wang and Wang, 1999, Fan and Wang, 2003, Buchler and Goldstein, 1999b]). Primary sequence and tertiary structure are nothing else than an informational channel [Shannon, 1948]: the information is encoded in the primary sequence and is transmitted through the secondary structure to the tertiary structure. Calling $h_{\rm seq}$, $h_{\rm sec}$, $h_{\rm ter}$ the informations of the three channel steps, one must have

$$h_{\rm seq} > h_{\rm sec} > h_{\rm ter} \tag{3.1}$$

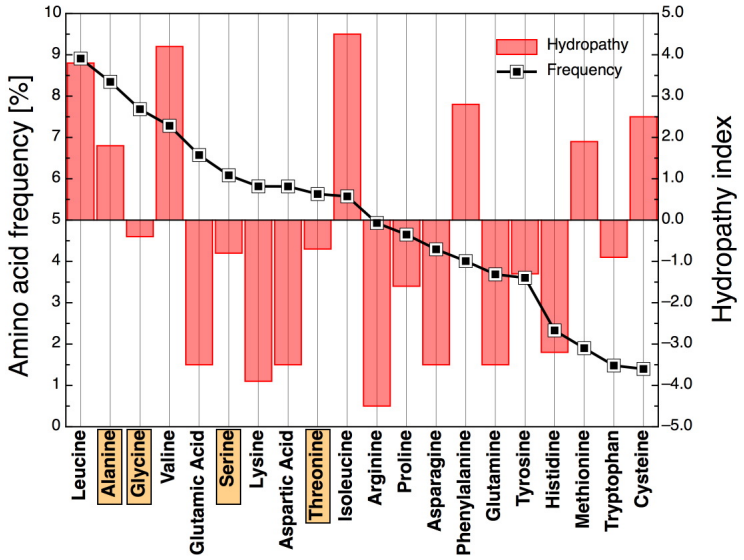

Figure 3.1: Amino acid propensities (left y axis and black curve) from the dataset found in [Dayalan et al., 2006]. Hydropathy index of the amino acids (right y axes and red bars) from [Kyte and Doolittle, 1982].

Sequence complexity can be evaluated from the Shannon entropy of sequence fragments of X-ray protein structures [Strait and Dewey, 1996] while the secondary structure and structural complexity can be estimated through the Shannon entropies respectively of the secondary structure strings SSS[8] and mesoscopic discrete rotational strings SRA[4] for the configurational space (see chapter 2). Structural complexity of X-ray structures has been estimated in a similar way by Levitt in [Park and Levitt, 1995]. To evaluate in first approximation the informations $h_{\rm seq}$, $h_{\rm sec}$, $h_{\rm ter}$ one can first assume equiprobable distributions of the amino acids so that the mean information per residue is $h_{\rm seq} = \ln_2 20 = 4.32$ bit/res as there are 20^R a priori possible R residue sequences (information is measured in bit when $\ln_2$ are used); according with the DSSP code [Andersen et al., 2002] (see chapter 2), secondary structure of a residue is encoded in a 8 letters alphabet so that assuming equi-probability one has $h_{\rm sec} = \ln_2 8 = 3$ bit/res; finally, assuming 4 dihedral states per residue one has $h_{\rm ter} = \ln_2 4 = 2$ bit/res. Thus the informations so estimated satisfy to the condition of the inequality 3.1. Taking into account the amino acid frequencies from X-ray databases is instructive. From the database PISCES protein sequence culling server (dataset cullpdb_pc90_res2.0_R0.25) a database called DASSD (Dihedral Angle and Secondary Structure Database of Short Amino acid Fragments) that contains dihedral angle values and secondary structure details of short amino acid fragments of lengths 1, 3 and 5 has been created by an australian bioinformatic group. Information stored in this database were extracted from a set of 5,227 non-redundant high resolution (less than 2-angstroms) protein structures [Dayalan et al., 2006]. From the data of the DASSD database we extracted frequencies of amino acids, secondary structure and dihedrals. With the amino acid frequency values, shown in figure 3.1, the information per residue is $h_{seq} = -\sum_{\rm i=1}^{20} p_i \ln_2 p_i \approx 4.1$ bit/res, which turns out not to be very different from the flat case. Interestingly the same result we obtained estimating the information on the non-redundant set of protein sequences that can be found

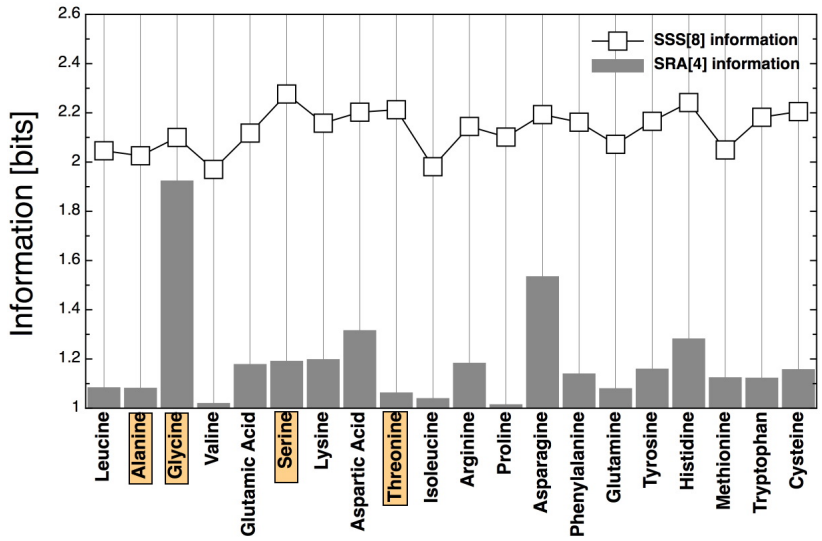

Figure 3.2: Amino acid informations for secondary structure SSS[8] (empty squared) and mesoscopic dihedral states SRA[4] (gray bars) estimated from the dataset found in [Dayalan et al., 2006]

on the NCBI website (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz). This database contains several millions of protein sequences from all known organisms. Frequencies of the most important elements of secondary structure per residue from the DASSD database result to be: 2% B (Isolated bridge), 17% coil, 23 % E (β-extended), 19 % T (turn-loop), 4 % G (3_{10}-helix) and 35 % H (α-helix). With these values of global propensities one obtains the mean information $h_{\rm sec} = 2.2$ bit/res. Finally the global statistics of mesoscopic dihedral states SRA[4] gives 44 % for 0, 51 % for 1, 2.4 % for 2 and 2.6 % for 3 (see figure 2.2 for the definitions of the mesoscopic states), which finally give a mean information $h_{\rm ter} = 1.3$ bit/res. Information per residue type for both secondary structure and dihedral mesostates are shown in figure 3.2. The informations per residue measured in bits can be related to an effective alphabet length through the relation $m_{\rm eff} = 2^h$. For the informations so far estimated one obtains respectively $2^{h_{\rm seq}} \sim 18$, $2^{h_{\rm sec}} \sim 4.6$ and $2^{h_{\rm ter}} \sim 2.5$. As pointed out in [Shakhnovich, 1998] a necessary condition to protein like sequences, namely sequences showing an energy gap between folded and unfolded state, is that the number of effective of amino acids $m_{\rm eff} > \gamma$ where γ represents the effective number of conformations per residue. This condition is satisfied when the conformers per residue are assumed as mesoscopic rotational state. We thus postulate that this condition may in principle be forced to hold by using a reduced residue alphabet, which in turns would have the counter effect of obtaining lower energy gaps between a folded and unfolded state (a lower thermodynamic control). Conversely we make the hypothesis that such alphabet reduction assure an higher kinetic accessibility to a designed folded state.

With all the previous premises in mind we propose a simplifying sequence strategy with the purpose to consider a three letter amino acid alphabet. In our simplified scheme we assume the secondary structure propensities of the amino acids as a drive-line for the choice of the amino acids forming the reduced alphabet. In particular we choose four amino acid types on the basis of their secondary structure properties: Alanine for helices, Glycine and Serine for turns and Threonine for β-extended. The table 3.2 resumes the secondary structure and SRA[4] propensities estimated from the DASSD database. With these

Residue	Coil [%]	E [%]	H+G+I [%]	T [%]	0 [%]	1 [%]	2 [%]	3 [%]
ALA	13	18	53	16	35	62	2	1
GLY	30	15	32	37	27	25	13	35
SER	22	21	30	25	49	48	2	1
THR	21	30	28	20	54	45	0.5	0.5

Table 3.1: Secondary structure and mesoscopic dihedral propensities estimated using the DASSD database [Dayalan et al., 2006].

residue types we constructed two kind of toy sequences: GST and AGT sequences, the former used for toy β-sheet mini proteins and the latter for a simplified version of the 56-residue B1 immunoglubulin-binding of streptococcal protein G (pdb code 1pgb) [Gallagher et al., 1994]. In this selection of amino acid threonine plays a special role as it is only weakly more β-prone than helix-prone. Furthermore its amphiphilic character (figure 3.1) together to the fact that is an uncharged, make of it a valid amino acid candidate to be simulated with the SASA implicit solvent [Ferrara et al., 2002], which has been not modeled to accurately describe charged residues. Thus for its nature, double secondary structure character and amphiphilicity, we postulate threonine shall not to be too sensitive to the inprecisions of both the force field and the solvation model. Finally we expected that toy sequences constructed with threonine should lead to a sort of "liquid" configurational spaces, in which diffusions among very different conformational basins is fast.

3.3 Methods

3.3.1 Sequences

Using the GST residues we constructed toy sequences in such a way their folded state could resemble a β-sheet motif. In particular 4 sequences have been considered with GS residues having the role of β-turns and T to favor the β-sheet formation. Respectively we constructed a 20-, 28-, 36-, 44- residues which are composed by polyTHR stretches linked by respectively 2, 3, 4, 5 glycine-serine turn-loops. GST-sequences are a sort of simplified-extended version of the GSGS peptide which we have studied in chapter 2. The sequences are constructed according with the SSS[8] string of the GSGS folded state (see table 3.2), in particular, where the secondary structure result with the letters "E" (extended) and "SS" (β-turn) respectively a THR and GLY-SER residues are replaced in the original GSGS sequence. Longer GST sequences are constructed extending the sequence template $T_5GST_6GS \cdots GST_6GST_5$. We called such sequences polyTHR_xGS with x the number of GS loops linking the polyTHR stretches. With such a template these sequences can be considered as repeat proteins. The sequence identity between the GSGS and polyTHR_2GS is 40 %. A similar procedure is followed to construct the simplified sequence of protein G in which AGT residues are used. We took the X-ray structure 1pgb and computed its secondary structure string (see table 3.2) which is used as reference for the simplifying scheme. The secondary structure of protein G written in a modular way is $Strand_1$-$Loop_1$-$Strand_2$-$Loop_2$-Helix-$Loop_3$-$Strand_3$-$Loop_4$-$Strand_4$. Wild type sequences are modified as follow: THRs replace strand residues,

GSGS	20-res	**T**WIQN**GST**KWYQN**GST**KIY**T**
SSS[8]		-EEEESSEEEEEESSEEEE-
SRA[4]		000021000000210000
polyTHR-2GS	20-res	TTTTTGSTTTTTTGSTTTTT
polyTHR-3GS	28-res	TTTTTGSTTTTTTGSTTTTTTTGSTTTTT
polyTHR-4GS	36-res	TTTTTGSTTTTTTGSTTTTTTTGSTTTTTTTGSTTTTT
polyTHR-5GS	44-res	TTTTTGSTTTTTTGSTTTTTTTGSTTTTTTTGSTTTTTTTGSTTTTT
1pgb	56-res	**TT**YKLVINGKTLKGE**TTT**KAVD**A**ET**A**EK**A**FKQY**A**NDN**G**VD**G**VW**T**YDDATK**T**FTV**T**E
SSS[8]		-EEEEEEEESSEEEEEEE-SSHHHHHHHHHHHHHHH----EEEEETTT-EEEEE-
SRA[4]		00000000110000000001011111111111111113000000000111300000
1pgb_AGT	56-res	TTTTTTTTTGGTTTTTTTTTGGAAAAAAAAAAAAAAAGGTTTTTTTTGGTTTTTTT

Table 3.2: The simplified protein sequences here studied: polyTHRs and 1pgb_AGT. GSGS and 1pgb sequences are reported with the secondary structure and SRA[4] strings that correspond to the simulation structure and the X-ray (see [Gallagher et al., 1994] for the latter) respectively.

GLYs replace loop residues and ALAs replace the helix residues. For simplicity the loops were made with double GLYs (other schemes have been tried: for example $loop_3$ with a 4-residue polyGLY but the data are not shown here). Thus we have a polyTHR from residue 1 to residue 9, double GLY at 10 and 11, polyTHR from 12 to 20, double GLY at 21 and 22, polyALA from 23 to 37, double GLY at 38 and 39, polyTHR from 40 to 47, double GLY at 48 and 49 and finally polyTHR from 50 to 56. The sequence so constructed is moderately hydrophobic due to the polyALA stretch. To construct a model structure, which is used for structural comparison, all the residue substitutions were made on the 1pgb structure by using the program Swiss Prot viewer (http://www.expasy.org), successively the structure was minimized with the program CHARMM [Brooks et al., 1983]. The sequence identity between 1pgb and 1pgb_AGT is 25%.

3.3.2 Molecular dynamics simulations

The simulations and part of the analysis of the trajectories were performed with the program CHARMM [Brooks et al., 1983]. PolyTHR_xGS and 1pgb_AGT polypeptides were modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field [Brooks et al., 1983]). As previously mentioned a mean field approximation based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent on the solute [Ferrara et al., 2002]. The same force field and implicit solvent model have been used recently in molecular dynamics simulations of the early steps of ordered aggregation [Gsponer et al., 2003, Cecchini et al., 2004, Cecchini et al., 2006] and folding of structured peptides (α-helices and β-sheets) ranging in size from 15 to 31 residues [Ferrara et al., 2000, Ferrara et al., 2002, Hiltpold et al., 2000] as well as small proteins of about 60 residues [Gsponer and Caflisch, 2001, Gsponer and Caflisch, 2002]. The force field alone resulted also very effective in the mesoscopic treatment of protein aggregation [Pellarin and Caflisch, 2006, Pellarin et al., 2007]. Despite the absence of collisions with water molecules, in the simulations with implicit solvent the separation of time-scales is comparable with that observed exper-

Protein	M	N (SSS[8])	N (SRA[4])	N (Cα-RMSD[5.0] clustering)
polyTHR_2GS	10^6	156314	356712	/
polyTHR_3GS	10^6	/	417909	/
polyTHR_4GS	10^6	/	589907	/
polyTHR_5GS	10^6	/	760681	/
1pgb_AGT 330 K	$5.3 \cdot 10^5$	519881	472630	132006
1pgb_AGT 300 K	$2.9 \cdot 10^5$	/	193820	2058

Table 3.3: The total number of sampled microstates M in the simulations; the total number N of mesoscopic strings SSS[8], SRA[4] and the total number of clusters found.

imentally. Helices fold in about 1 ns, β-hairpins in about 10 ns [Ferrara et al., 2000] and triple-stranded β-sheets in about 100 ns [Ferrara and Caflisch, 2000] while the experimental values are $\sim$0.1 μs, $\sim$1 μs [Eaton et al., 2000] and $\sim$10 μs [De Alba et al., 1999] respectively.
For each of the polyTHR_xGS toy proteins a set of 4 independent molecular dynamics simulations were performed, 5 μs long each for a total length of 20 μs. The total number of collected microstates is 10^6 snapshots for each sequence with a saving time step of 20 ps. The simulation temperature was set to 330 K and the starting conformation for all the simulations was a completely extended structure. Regarding the 1pgb_AGT protein a set 4 independent molecular dynamics simulations were also performed, ~ 2.5 μs long each for a total length of about 10 μs. The simulation temperature was set to 330 K and the starting conformation for all the simulations was yet a completely extended structure. For this sequence also 3 independent stability simulations at 300 K were performed in which the initial structure was the 1pgb X-ray with the simplified sequence. These stability simulations are 2 μs long each for total simulation time of 6 μs. The total number of collected microstates are $\sim 5 \cdot 10^5$ and $\sim 3 \cdot 10^5$ respectively for the simulations at the temperatures 330 and 300 K.

3.3.3 Description of the configurational space

All the time series of trajectories obtained from the molecular dynamics simulations were mapped into new time series of mesoscopic rotational strings. For the sequences polyTHR_GS the description SRA[4] was adopted (see chapter 2). The total number of strings obtained from the mapping procedure are reported in table 3.3.2. For the simulations on 1pgb_AGT the trajectories were mapped into time series of both SSS[8] and SRA[4] mesoscopic strings. Moreover a cluster analysis on the whole ensemble of microstates was performed using the Cα-RMSD based program CLUSTER which has been described previously in chapter 2. Many RMSD cutoffs were tried out to clusterize the sampled conformation spaces of 1pgb_AGT protein, these trials eventually lead to a cutoff of 5 Å. This cutoff was particularly effective in capturing the salient structural motifs of the stable states of 1pgb_AGT because it revealed to be a good compromise between coarse-grained accuracy and computational clustering efficiency. The total number of clusters obtained are reported in table 3.3.2. Thus, microscopic trajectories of 1pgb_AGT were mapped in new time series of clusters for the analysis.

3.4 Results

3.4.1 Statistical thermodynamics of the configurational spaces

From the time series of the SRA[4] mesostates the thermodynamic quantities can be estimated according to the methods we have presented in chapter 2. In particular from a trajectories of M microstates N mesostates are extracted so that the following quantities are estimated: the probability of the mesostate P_i, the mean effective energy (potential plus solvation energy) difference per mesostate $\Delta E_i = \overline{E_i} - \overline{E}$, with $\overline{E}$ the mean total effective energy and $\overline{E_i}$ the mean effective energy per mesostate; the mean internal entropy difference per mesostate $T\Delta S_i^b = T(S_i^{\mathrm{b}} - S^{\mathrm{b}})$ where

$$S^{\mathrm{b}} = \sum_{i=1}^{N} P_i S_i^{\mathrm{b}} \tag{3.2}$$

is the mean total internal entropy of the ensemble of mesostates with

$$S_i^{\mathrm{b}} = 1/2k_B(\beta\sigma(E_i))^2 \tag{3.3}$$

the internal entropy per mesostate; the mean free energy difference per mesostate is thus estimated as

$$\Delta G_i = \Delta E_i - T\Delta S_i^{\mathrm{b}} \tag{3.4}$$

Mesostates with $\Delta G_i < 0$ represent statistically stable mesostates (well sampled mesostates), while those having $\Delta G_i \sim 0$ are fluctuations and are gaussian distributed (mesostates which are indistinguishable from microstates). Mesostates that have a negative configurational entropy loss are enthalpy driven (their free energy is favorite by their low mean effective energy $\overline{E_i}$) while those that have a positive entropy loss are favored by the internal entropy S_i^{b}. Moreover, we estimated the configurational entropy loss from the Shannon entropy of the ensemble of mesoscopic strings. The letter probabilities per string site $p_s(0), p_s(1), p_s(2), p_s(3)$ corresponding to SRA[4] are estimated from the time series of the strings so that the total configurational entropy is given as

$$h = \sum_{s=1}^{R} h_s \tag{3.5}$$

with

$$h_s = -k_B \sum_{k=0}^{3} p_s(k) \ln p_s(k) \tag{3.6}$$

with R the string length. Thus for each string $\mathbb{S}_i$ one estimates the configurational entropy loss such as $\Delta h(\mathbb{S}_i) = h(\mathbb{S}_i) - h$ (see section 2.5.1).

Analysis of the polyTHR_xGS proteins

In table 3.4.1 the most populated 10 mesostates are shown with their corresponding mesoscopic string and thermodynamic quantities for the polyTHR_xGS proteins. The most populated mesostates correspond to a β-strand structure in all the cases which have a negative values of the configurational entropy loss, meaning that they are favored mainly by enthalpy. These β-strand structures represent the folded state for these proteins (see figure 3.3 for their ensemble representation). Conversely helix-like

i	ΔE_i $^{(*)}$	$\sigma(E_i)$ $^{(*)}$	P_i	$T\Delta h_i$ $^{(*)}$	$T\Delta S_i^{\mathrm{b}}$ $^{(*)}$	ΔG_i $^{(*)}$	Mesostate
polyTHR_2GS							
1	-4.8	8.4	0.143	-2.7	7.6	-12.4	**000021000000210000**
2	-3.3	9.7	0.019	2.3	24.7	-28.0	111111111111111111
3	-4.9	8.4	0.011	-2.2	7.2	-12.1	000021000001110000
4	-4.1	8.7	0.010	-2.2	11.5	-15.6	000021000000210010
5	-4.9	8.4	0.010	-2.1	7.4	-12.3	000111000000210000
6	-3.0	9.5	0.007	1.6	22.7	-25.7	111111111111111110
7	-2.2	9.2	0.007	-2.0	17.8	-20.0	000021000000200000
8	-2.8	9.6	0.005	1.8	23.7	-26.6	011111111111111111
9	-4.7	8.6	0.005	-0.9	10.2	-14.9	110010000000210000
10	-3.9	8.8	0.005	-2.2	12.6	-16.5	100021000000210000
polyTHR_3GS							
1	-7.8	9.4	0.062	-3.6	40.8	-48.7	**000021000000210000000210000**
2	-7.7	9.3	0.013	-3.1	40.2	-47.9	000021000000210000000210010
3	-2.5	10.2	0.009	2.7	53.0	-55.5	111111111111111111111111111
4	-7.8	9.4	0.008	-3.0	40.9	-48.8	000111000000210000000210000
5	-7.7	9.3	0.007	-3.4	40.1	-47.9	000021000001110000000210000
6	-9.1	9.4	0.005	0.3	41.9	-51.0	000000111111300000000210000
7	-2.2	10.2	0.004	1.9	53.7	-55.8	111111111111111111111111110
8	-4.5	9.7	0.003	-3.0	46.3	-50.7	000010000000210000000210000
9	-5.7	9.8	0.003	-3.0	47.7	-53.4	000021000000210000000200000
10	-7.3	9.3	0.003	-3.0	40.2	-47.6	100021000000210000000210000
polyTHR_4GS							
1	-8.3	11.6	0.036	-5.9	82.7	-91.0	**000021000000210000002100000021000000210000**
2	-8.6	11.5	0.009	-5.2	80.5	-89.2	000021000000210000002100000021000000210010
3	-8.5	11.4	0.008	-5.5	80.3	-88.8	000021000001110000002100000021000000210000
4	-8.1	11.6	0.004	-5.2	83.9	-92.0	000111000000210000002100000021000000210000
5	-7.1	11.9	0.002	-5.2	89.0	-96.2	010011000000210000002100000021000000210000
6	-8.0	11.4	0.002	-5.2	80.0	-88.0	000021000000210000011100000021000000210000
7	-8.7	13.7	0.002	6.6	123.5	-132.3	111111111111111111111111111111111111
8	-4.2	11.6	0.002	-5.3	83.4	-87.6	000010000000210000002100000021000000210000
9	-7.0	12.2	0.002	-5.2	94.7	-101.7	100021000000210000002100000021000000210000
10	-4.9	12.8	0.002	-5.4	104.6	-109.5	000021000000210000002100000021000000200000
polyTHR_5GS							
1	-11.7	12.1	0.0154	-6.7	103.7	-115.4	**000021000000210000002100000021000000210000**
2	-12.2	12.0	0.0034	-6.1	102.1	-114.3	000021000000210000002100000021000000210010
3	-11.6	12.1	0.0031	-6.4	103.1	-114.8	000021000001110000002100000021000000210000
4	-12.2	12.0	0.0021	-6.3	101.8	-114.0	000021000000210000011100000021000000210000
5	-11.7	12.2	0.0017	-6.1	106.2	-117.9	000111000000210000002100000021000000210000
6	-18.8	12.8	0.0013	-1.6	116.2	-135.1	000130010111300010002100000021000000210000
7	-5.9	12.5	0.0012	-4.3	112.0	-117.9	000000000000210000000000000000021000000000000
8	-11.1	11.6	0.0008	-6.3	95.2	-106.3	000021000000210000002100000111000000210000
9	-9.4	12.5	0.0008	-6.5	110.9	-120.3	000021000000210000002100000021000000200000
10	-4.1	13.2	0.0008	-4.6	123.8	-127.9	000010000000210000000000000021000000000000

Table 3.4: $^{(*)}$ [kcal/mol]. The list of the first 10 most populated SRA[4] mesostates from the polyTHR_xGS simulations with the thermodynamic quantities estimated according to the equations developed in chapter 2. In particular ΔE_i is the effective mean energy difference per mesostate, $\sigma(E_i)$ is the standard deviation of $\overline{E_i}$, P_i is the population per mesostate, $T\Delta h_i$ is the configurational entropy loss per mesostate, $T\Delta S_i^{\mathrm{b}}$ is the internal entropy difference per mesostate and ΔG_i is the free energy difference per mesostate.

mesostates ("1" rich strings) have a positive entropy loss which means that they are mainly stabilized by their internal entropy. On this respect, as it is shown later, β-strand and helix-like basins are thermodynamically competitive states as their configurational entropy has opposite sign. In figure 3.3 (A) we show the ensemble view of the most populated mesostates for the polyTHR_xGS proteins, respectively three-, four-, five-, six-stranded β-sheets. From the ensemble of microstates which are member of the folded mesostate mean structures were computed. these structure were used to calculate the time series of the Cα-RMSD from the polyTHR_xGS simulations. In figure 3.4 these time series are shown. All the four time series clearly indicate that the polyTHR_xGS proteins reversibly fold at the 330 K. In particular many folding events are observed for all four the proteins, they result to be about 250 for polyTHR_2GS, 110 for polyTHR_3GS, 40 for polyTHR_4GS and 20 for polyTHR_5GS. From the frequencies per residues

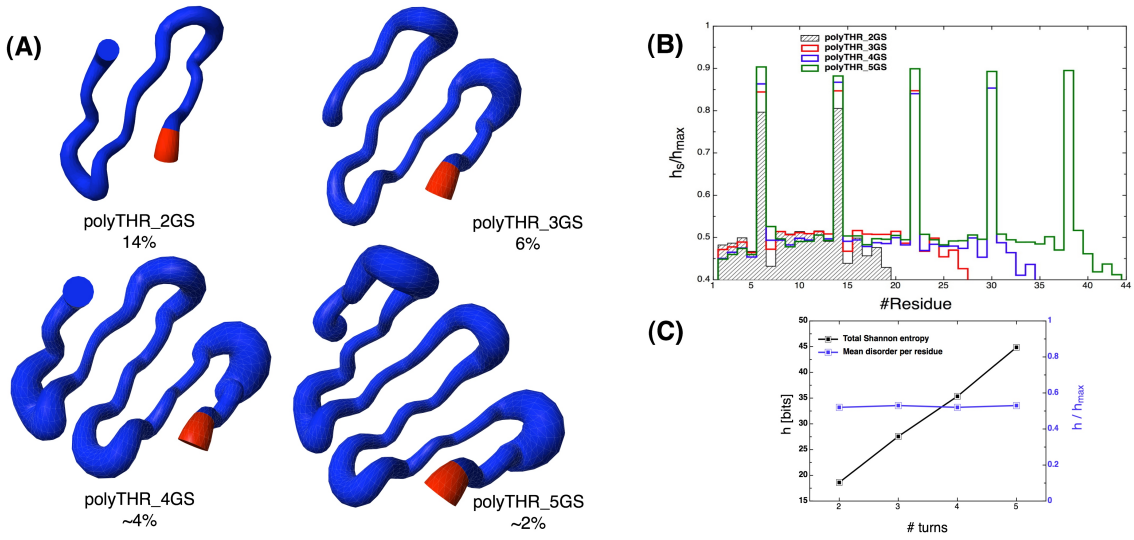

Figure 3.3: (A) The ensemble representation of the most populated folded mesostates for the polyTHR_xGS, respectively three-, four-, five-, six-stranded β-sheets. These mesostates are mainly promoted by the low enthalpy. (B) The normalized Shannon entropy per residue (also called disorder) from the ensemble of SRA[4] mesostates obtained from the polyTHR simulations. High disorder peaks correspond to the Gly residues at the turn positions. Disorder profiles look very similar among the sequences showing a modular pattern. (C) The total Shannon entropy (black curve) and the mean disorders per residue respectively as functions of the number of GS turns. The disorder does not depend on the protein size while the total entropy linearly increases with the number of GS turns, that is the chain size.

of the four rotational states SRA[4] the normalized Shannon entropy (as introduced in chapter 2 we call it statistical disorder) per residue is calculated by using the relation $h_s = -\sum_{k=0}^{3} p_s(k) \ln p_s(k) / \ln 4$ where s is the residue number. The disorders per residues are shown in figure 3.3 (B). Peaks correspond to the Gly residues and the overall profiles are similar among the different polyTHR proteins suggesting a modular structure of the configurational space. In figure 3.3 (C) the Shannon entropy and the mean disorder per residues as a function of the number of GS turns are shown. The total Shannon entropy increases linearly with the number of GS turns while the mean disorder per residue is independent on the protein size. The former feature essentially depends on the extensivity of the entropy which naturally increases with the system size, while the disorder, which is an intensive quantity tell us that the configurational space per residue is insensitive of the system size. This means that these "designed" proteins, made up of repeated units posses a configurational space shaped almost exclusively by local interactions. Local interactions in this case are somehow biased by the secondary structure propensities of the residues we used for the sequence design. Each strand interacts with its contiguous strand with the turns playing the role of mechanical joints. The mean value of the disorder per residue is 0.53 which corresponds to an entropy per residue of about 1 bit which means that in average each residue admits two main states: helix-like and β-extended.

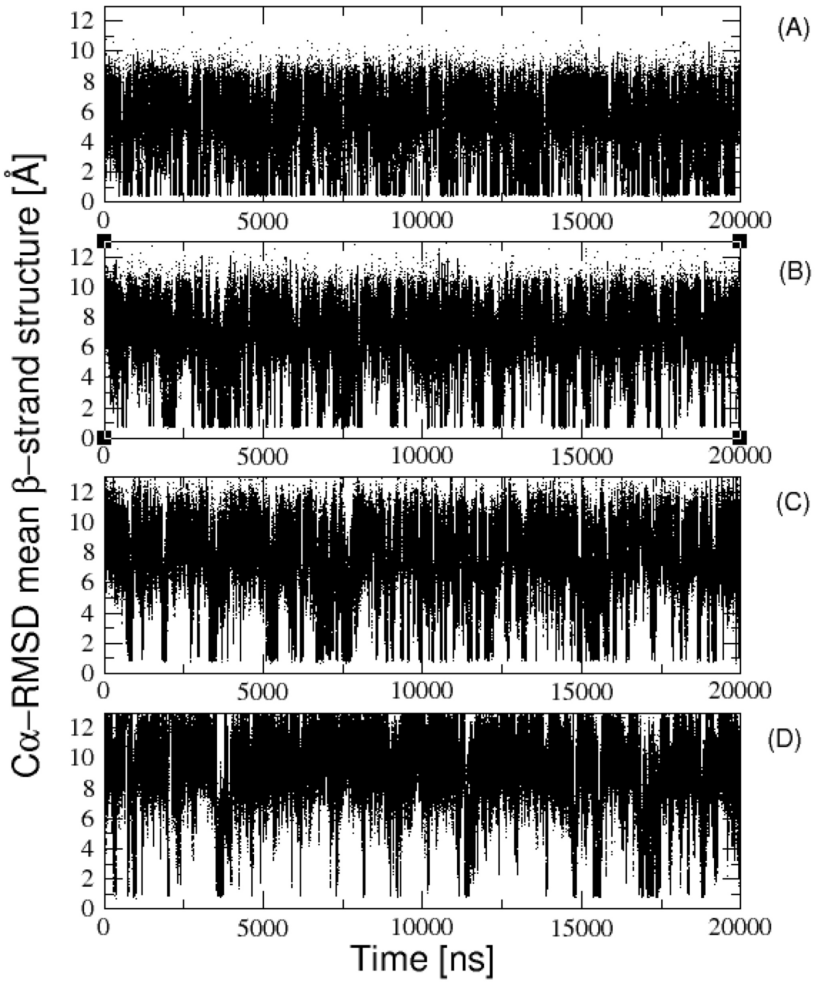

Figure 3.4: Time series Cα-RMSD of the polyTHR_xGS folding simulations at 330 K with respect to the averaged β-strand structures. The model mean structures are computed from the ensemble of structures which are members of the most populated SRA[4] mesostate. Many spontaneous folding events are observed for all four the proteins, they result to be about 250 for polyTHR_2GS (A), 110 for polyTHR_3GS (B), 40 for polyTHR_4GS (C) and 20 for polyTHR_5GS (D).

Analysis of the 1pgb_AGT protein

Two kind of descriptions were used to characterize the configurational space of the protein 1pgb_AGT: one based on string of mesostates SRA[4] and one based on Cα-RMSD clustering. The former description is a truly discretization of the dihedral configurational space which is not coarse enough to produce highly populated mesostates. In fact the number of string of mesostates is very close to the total number of microstates. Nevertheless such description is very useful to investigate the organization of the configurational space of 1pgb_AGT. In table 3.6 the list of the first 50 most populated SRA[4] strings are reported together with their thermodynamical quantities. Strings in bold are those which corresponds

to the α/β topology of G protein 1pgb. For these mesostates the mean effective energy difference is very favorable with respect to the mean total mean effective energy ($\Delta E_i \sim -22$ kcal/mol) and correspondingly the internal entropy differences are in averages less favorable with respect to other mesostate. This means that these mesostates are definitely promoted by the enthalpy as in fact can be seen from the values of the configurational entropy loss $T\Delta h_i$, which is for those mesostates always negative (about -6 kcal/mol). The list of SRA[4] mesostates does not tell us what is the most populated structural basin which can be assumed as folded state for the 1pgb_AGT protein. On this respect the Cα-RMSD clustering at 5Å solves the problem of finding the folded state (for the clustering procedure the program CLUSTER was used whose features are explained in chapter 2). In table 3.6 are listed the first most populated clusters as well as with their thermodynamic parameters and structural representations. The total number of clusters obtained is about 132000 on a set of about 530000 conformers. Among the clusters only 3124 have ≥ 2 structures. The most populated cluster has the 3.3 % of the total weight. To this cluster corresponds the structural topology of the G protein 1pgb. Many other cluster share the same correct topology, for instance clusters 4, 10, 11, 17 in table. Many clusters are characterized by having the correct native secondary structure but an incorrect native topology. This can be realized with the four stranded β-sheet with either one or two misplaced hairpins. Examples of such clusters are the 2, 3, 9, 13, 14, 18 shown in figure. From these structures the full folded state is reachable by disrupting a hairpin and inverting the orientation of a β-turn. As it will shown later these states are in a fast equilibrium with the full correct folded state. That already suggest that the folding state for this toy protein resembles that of a molten globular protein in which the secondary structure can be well defined while the structural definition can be shallow. Interesting are also clusters having the β strand well formed but the α-helix either coil of β. It is not easy for this states reaching the full correct folded states in few steps. Finally kinetic traps fiber-like are also sampled, these are completely off pathway states. An example of unfolded clusters are the cluster 8, 16 and 19 in which helices involve Thr residues and are close packed with an either double or three stranded β-sheet. From the X-ray structure 1pgb [Gallagher et al., 1994] time series of Cα-RMSD were computed for both the simulations at 300 and 330 K. Simulations at 300 K were started from the X-ray topology (namely the structure having the simplified AGT sequence and the X-ray backbone) to test how stable was the protein G topology in the simplified 1pgb_AGT protein. In figure 3.5 (A) the Cα-RMSD time series with respect the X-ray topology are shown for these three independent simulations. The lag time before unfolding is about 500 ns and it can be assumed as a Poisson event. In one of the three simulations at 300 K the protein is unfolded at $\sim$10 Å of Cα-RMSD and refolds. The Cα-RMSD time series at 330 K (figure 3.5 (B)) indicates that the protein folds reversibly and a total number of about 15 folding events can be estimated. That is a remarkable result considering that all the simulations at 330 K were started from a completely extended conformation so that the protein spontaneously can find its folded state which in particular coincides with that of the protein 1pgb. From the Cα-RMSD time series at both the temperatures the mean effective energy $E_{\rm eff}$ for Cα-RMSD ranges of 0.5 Å were computed starting from RMSD=2 Å. In figure 3.5 (C) the mean effective energy differences $\Delta E_{\rm eff}(\mathrm{RMSD}) = \langle E_{\rm eff}(\mathrm{RMSD})\rangle - \langle E_{\rm eff}(\mathrm{RMSD} \leq 2)\rangle$ are plotted as a function of the Cα-RMSD. All the ensemble of structures having RMSD$\leq$ 3 Å are in an absolute minima of the effective energy, in particular the enthalpy loss from from RMSD$\gtrsim$ 8 Å to RMSD$\leqslant$ 2 is about 20 kcal/mol at 330 K and about 9 kcal/mol at 300. This energy plot provides support to the hypothesis that the folded state of the 1pgb_AGT protein is at bottom of a free energy funnel. Later on this chapter further evidences will

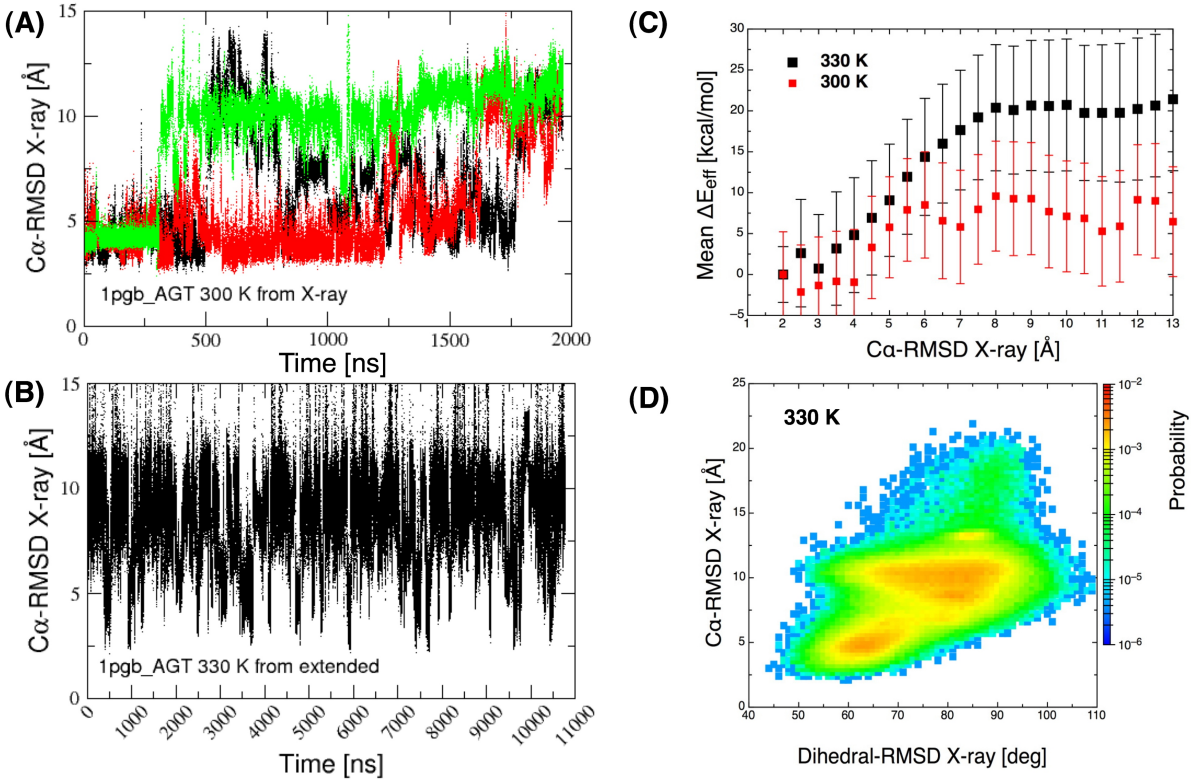

Figure 3.5: Time series Cα-RMSD of the 1pgb_AGT for the stability simulations at 300 K (A) and folding simulations at 330 K (B) with respect to the X-ray structure 1pgb. The firsts and lasts 2 Cα and the Gly residues were excluded from the RMSD calculations. In the time series at 330 K approximatively 30 folding events can be identified. In the time series at 300 K the folded state in maintained in its topology for about 500 ns and an unfolding event can be assumed as a Poissonian event. (C) The mean effective energy was computed on Cα-RMSD ranges of 0.5 Å for both the simulations performed at 300 and 330 K. On plot (C) the effective mean energy differences $\Delta E_{\text{eff}}(\text{RMSD}) = \langle E_{\text{eff}}(\text{RMSD})\rangle - \langle E_{\text{eff}}(\text{RMSD} \leq 2)\rangle$ as a function of the Cα-RMSD with respect to the X-ray structure. The curves clearly indicates that the protein G topology is the enthalpic minima for protein 1pgb_AGT at both the simulation temperatures. An enthalpy difference can be estimated between folded and unfolded: approximatively 19 kcal/mol at 330 K and 9 kcal/mol at 300 K. In (D) a two dimensional density plot of the Cα-RMSD as a function of the Dihedral-RMSD, both with respect to the X-ray structure 1pgb. All the residues were taken into account. The plot clearly shows the presence of two broad phases, one ranging from 2.5 to 6 Å and from 55 to 70 deg in Dihedral-RMSD and the other much broader. Within the first phase the folded state can be located. Unlike the Cα-RMSD the Dihedral-RMSD is more sensitive to the local similarities to the X-ray structure. Interestingly, in the broader phase the Dihedral-RMSD includes native like values ($\sim$60 deg) suggesting that in the unfolded state the elements of secondary structure can be locally already shaped.

i	ΔE_i $^{(*)}$	$\sigma(E_i)$ $^{(*)}$	P_i	$T\Delta h_i$ $^{(*)}$	$T\Delta S_i^{\mathrm{b}}$ $^{(*)}$	ΔG_i $^{(*)}$	Mesostate
1	-4.5	11.65	0.00050	4.9	96.29	-100.81	00000000000000000011000000000010000021000000000000000
2	-7.2	11.62	0.00020	5.1	95.79	-102.95	00000000000000000011000001000000000021000000000000000
3	-3.8	10.85	0.00019	4.9	82.71	-86.54	00000000000000000011000000100000000021000000000000000
4	-4.4	12.03	0.00019	5.1	103.11	-107.47	00000000000000000011000000000000010002100000000000000
5	-25.2	11.37	0.00015	-6.1	91.34	-116.55	**10000000021000000001211111111111111130000010102000000**
6	-3.6	12.16	0.00013	4.9	105.56	-109.18	00000000000000000011000000001000000021000000000000000
7	-1.3	12.58	0.00012	6.5	113.30	-114.61	00000000000000000100130000100000000021000000000000000
8	-3.9	10.54	0.00011	6.5	77.60	-81.52	00000000000000011001300000000001000021000000000000000
9	-4.1	12.35	0.00011	5.1	108.95	-113.03	00000000000000010011000000000001000021000000000000000
10	-15.0	13.47	0.00011	-4.3	130.96	-145.96	**00010000021000000002000111111111100110000000021000000**
11	-8.2	12.27	0.00009	6.4	107.46	-115.62	00000002100000000200010010000000021000000001001001
12	-8.2	13.10	0.00009	5.0	123.52	-131.76	00000000000000000010000000000010000021000000000000000
13	-2.3	11.66	0.00008	5.2	96.44	-98.70	00000000000000001001000000000001000021000000000000000
14	-0.5	12.04	0.00008	6.3	103.23	-103.69	00000000000000001001300000000001000021000000000000000
15	-2.1	11.87	0.00008	5.5	100.17	-102.25	00000000000000000012000000000001000021000000000000000
16	-1.1	11.08	0.00008	5.2	86.55	-87.66	00000000000000001001000001000000000021000000000000000
17	-17.0	13.66	0.00008	-4.2	134.92	-151.88	**10000000021000000001211111111110110000100011000000010**
18	-22.0	13.33	0.00008	-5.9	128.18	-150.16	**10000000020000000013011111111111111130000000020000000**
19	-4.3	7.69	0.00008	6.7	38.30	-42.60	00000000000000011001300001000000000021000000000000000
20	-22.4	12.17	0.00007	-6.6	105.60	-128.03	**10000000020000100001211111111111111130000000020000000**
21	-6.3	10.77	0.00007	6.3	81.41	-87.66	00000000000000001001300000010000000021000000000000000
22	-3.7	10.33	0.00007	7.0	74.28	-78.01	00000000000000000013000000000000000021000000000000000
23	-9.3	12.03	0.00007	6.5	103.13	-112.40	00000000210000000020001001000000002100000000001001
24	-3.3	10.57	0.00007	6.1	78.14	-81.43	00000000000000000011000000000000000021000000000000000
25	-4.5	11.94	0.00007	5.2	101.48	-106.01	00000000000000011001000000000010000021000000000000000
26	-0.9	10.00	0.00007	6.1	69.26	-70.12	00000000000000001013000000000010000021000000000000000
27	-8.2	13.25	0.00006	6.1	126.56	-134.78	00000000000000000113000000000010000021000000000000000
28	2.0	7.70	0.00006	5.0	38.40	-36.37	00000000000000001001000000000010000021000000000000000
29	-18.9	11.23	0.00006	5.4	89.01	-107.86	10000000021000000002100000000010000110000010102000000
30	-24.4	12.80	0.00006	-6.5	117.64	-142.00	**10000000020000000013011111111111111100000000020000000**
31	-20.2	9.40	0.00006	-6.7	60.41	-80.66	**10000000021000000001211111111111111100000010102000000**
32	1.0	11.93	0.00006	6.2	101.35	-100.32	00000000000000001001000000000000000021000000000000000
33	-12.8	10.32	0.00005	5.8	74.18	-86.97	00000000210000000011000100100000000021000000001001001
34	-2.5	15.54	0.00005	6.2	176.53	-179.08	00000000000000000010300000000001000021000000000000000
35	-17.4	10.94	0.00005	-7.4	84.19	-101.58	**00000000210010000001011111111111111130000000021000000**
36	-3.5	11.84	0.00005	4.2	99.67	-103.15	00000000210000000000211000010000000021000000021000000
37	4.6	13.66	0.00005	5.4	134.83	-130.18	00000000000000001001000000000000010002100000000000000
38	-22.1	11.55	0.00005	-6.2	94.46	-116.53	**10000000021000000001011111111111111130000010102000000**
39	-19.1	12.38	0.00005	-2.8	109.63	-128.71	00010000021000000000200011111111100021000000021000000
40	-4.7	11.43	0.00005	6.6	92.40	-97.12	00000000000000000012000000000000000021000000000000000
41	-36.8	14.81	0.00005	-3.3	159.55	-196.35	00000011000111111111111111111111111111101111111011100000
42	4.7	13.09	0.00005	6.4	123.35	-118.60	00000000220000000010000000000000000021000000000000000
43	-21.8	12.94	0.00005	-7.3	120.26	-142.06	**00000000210010000001211111111111111130000000021000000**
44	-20.3	13.30	0.00004	-6.8	127.59	-147.90	**00000000210010000002011111111111111130000000021000000**
45	-21.1	12.09	0.00004	-6.5	104.14	-125.25	00000000010111111111111111111111111113000000002100000
46	-8.0	11.32	0.00004	3.8	90.60	-98.55	00010100100000001000210010010000100001000000002100000
47	-15.3	14.59	0.00004	-6.3	154.66	-169.95	00000000121011111111111111111111111113000000002100000
48	-1.9	14.10	0.00004	5.3	144.17	-146.06	00000000000000001001100000100000000002100000000000000
49	-13.0	13.39	0.00004	6.0	129.37	-142.33	00000000210000000011000100100000000021000000000001001
50	-8.4	11.43	0.00004	5.7	92.44	-100.86	00000000210000000002000100100000000021000000000001000

Table 3.5: $^{(*)}$ [kcal/mol]. The list of the first 50 most populated SRA[4] mesostates from the 1pgb_AGT simulations at 330 K with their thermodynamic quantities. Strings in bold are those corresponding to a structural motif compatible with that of protein G. There are many strings sharing such a structural feature, notably these strings have a negative configurational entropy loss $T\Delta h_i$ which means they are promoted by enthalpy.

be given through the kinetic analysis of the space of mesostates. Other than the Cα-RMSD a dihedral backbone RMSD was computed with respect to the X-ray structure 1pgb. Given two backbone conformations C^0 and C with their dihedral angles $\{\phi_i^0, \psi_i^0\}$ and $\{\phi_i, \psi_i\}$ with i the dihedral pair index, the dihedral RMSD is defined as

$$\mathrm{diheRMSD}(C^0, C) = \sqrt{\frac{1}{2(R-1)} \sum_{i=1}^{R-1} ((\phi_i^0 - \phi_i)^2 + (\psi_i^0 - \psi_i)^2)} \pmod{180} \deg \tag{3.7}$$

where R is the number of residues of the protein. It is interesting to compare the dihedral RMSD with the Cα-RMSD: the former is more sensitive than the latter to the native-like secondary structure which

i	ΔE_i $^{(*)}$	$\sigma(E_i)$ $^{(*)}$	P_i	$T\Delta S_i^{\mathrm{b}}$ $^{(*)}$	ΔG_i $^{(*)}$	i	ΔE_i $^{(*)}$	$\sigma(E_i)$ $^{(*)}$	P_i	$T\Delta S_i^{\mathrm{b}}$ $^{(*)}$	ΔG_i $^{(*)}$
1	-13.9	14.1	0.033	69.0	-82.8	11	-8.7	13.9	0.007	64.0	-72.8
2	-10.8	13.7	0.019	60.4	-71.2	12	-5.6	13.6	0.004	57.2	-62.9
3	-11.0	13.6	0.014	57.6	-68.6	13	-8.3	13.6	0.004	58.6	-66.9
4	-8.6	13.8	0.012	61.8	-70.4	14	-10.4	13.8	0.004	62.2	-72.6
5	-9.9	13.8	0.008	62.2	-72.1	15	-10.5	14.1	0.004	69.5	-80.0
6	-2.6	12.7	0.008	39.7	-42.4	16	-15.7	16.5	0.003	124.5	-140.1
7	-4.1	13.7	0.008	59.9	-64.0	17	-8.1	13.9	0.003	63.5	-71.6
8	-18.4	15.2	0.008	92.4	-110.8	18	-5.8	14.0	0.003	65.7	-71.5
9	-7.8	14.1	0.007	68.7	-76.5	19	-12.1	14.5	0.003	78.2	-90.2
10	-14.8	13.9	0.007	65.1	-79.9	20	-7.1	13.7	0.003	60.5	-67.5

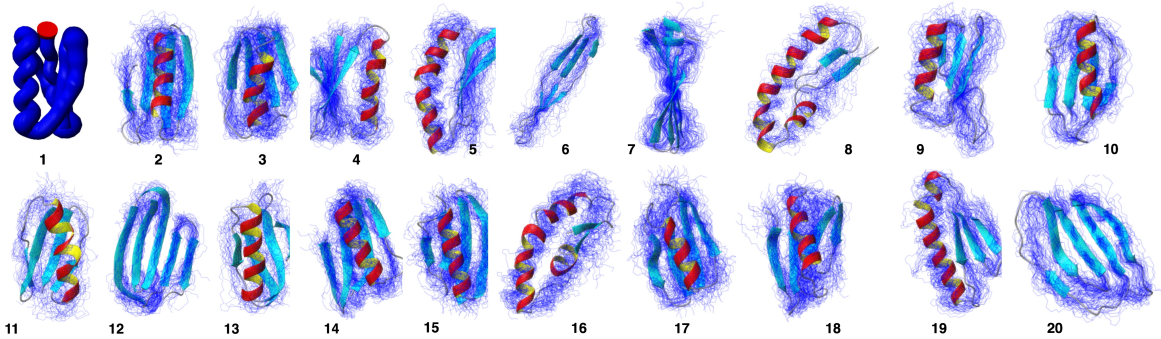

Table 3.6: $^{(*)}$ [kcal/mol]. The list of the first 20 most populated RMSD[5.0] clusters from the 1pgb_AGT simulations at 330 K with their thermodynamic quantities and structural representations. Many clusters correspond to the structural topology of protein G, in particular the clusters 1 (which is also shown in figure 3.6 (A) with its cluster center), 4, 10, 11, 17 and many others. There are also many clusters having the native secondary structure but an incorrect topology, for instance either one or two hairpins misplaced. These clusters are for example the 2, 3, 9, 13, 14, 18 among those shown in figure. Other interesting clusters are those having the β-sheet formed but the helix either in a coil or β structure: see for instance cluster 12 and 20.

can be locally shaped but globally misplaced with respect to a reference structure. On this respect a 2 dimensional density plot Cα-RMSD as function of dihedral RMSD was computed to project the conformational space on two basically different progress variable. This plot is shown in figure 3.5 and clearly reveals the presence of two main macrostates: the folded one centered at 3.5 Å of Cα-RMSD and 65 deg of dihedral RMSD, the second one centered at about 10 Å and 85 deg. Interestingly the second basin shows a wide range of variability in the dihedral RMSD, in particular for large values of Cα-RMSD native like values of the dihedral RMSD are quite probable. This indicates that although the protein can be unfolded the elements of secondary structure can as well as locally already formed. Such a condition can be realized for instance when the internal helix is formed along with one of the two hairpins. In this sense such an unfolded basin can be viewed as a sort of pre-folded state in which the elements of secondary structure diffuse within each other to assemble the folded state. Such a picture, at this stage of analysis already suggests that the folding mechanism of this toy protein might follow a diffusion-collisional model of folding [Karplus and Weaver, 1979].

The folded state of 1pgb_AGT corresponds to the topology of protein G. To investigate that more quantitatively all the microstates within the most populated folded cluster have been checked to share the same topology of the X-ray structure 1pgb. Moreover comparing the cluster center of the most popu-

lated cluster with the structure 1pgb a Cα-RMSD of 2.6 Å was obtained where the first and last two Cαs and Gly's Cαs were excluded. In figure 3.6 (A) the superposition of the cluster center and the structure 1pgb is shown. The packing between helix and β-sheet is surprisingly well reproduced although the structural detail is clearly not completely correct due to the fact that a simplified sequence was studied. In particular the hairpins are not completely overlapped because the lack of well defined hydrophobic core. All the structures (about 20000) which are members of the most populated cluster were compared with the structure 1pgb calculating the "autopair" Cα-RMSD. Such a method of structural alignment finds the best overlap between chain fragments that can also be non contiguous in sequence. It maximizes the number Cα pairs and minimize the Cα-RMSD between the chain fragments. In figure 3.6 (B) it is shown the result of this analysis, in particular for the number of Cα pairs adopted the mean and best Cα-RMSD were calculated. Interestingly a Cα-RMSD plateau within 50 and 85 % of the total pairs is revealed with a value not exceeding 2.6 Å. Moreover for 46 pairs a best value of 1.3 Å was found. These results provide quantitative evidence that the folded state for the protein 1pgb_AGT is strongly fluctuating. Notably the protein G overall topology can be fulfilled in a shallow manner, i.e. the β hairpins can be formed with a shift in the chain up to ± 2 residues. That is due to partial specificity of the toy protein sequence whose configurational space is determined solely by secondary contacts. This result supports the hypothesis that a simplified sequence whose amino acids are chosen according to their secondary structure properties, are able to encode molten globular protein structures, namely structures in which the folded secondary structure is well defined while tertiary contacts lack of specificity [Ptitsyn et al., 1990, Ptitsyn and Uversky, 1994, Vassilenko and Uversky, 2002].

3.4.2 Organization of the configurational space of "primitive" proteins

The description of the molecular dynamics simulations in terms of mesoscopic strings, notably either SRA[4] or SSS[4] strings, allows to effectively study the organization of the protein configurational spaces as extensively described in chapter 2. In figure 3.3 we noticed that the configurational entropy of the polyTHR_xGS polypeptides grows linearly with respect to the chain size. The normalized configurational entropy was computed for the protein 1pgb_AGT too. In figure 3.7 (A) are shown the normalized entropies respectively for the ensemble of SRA[4] string of mesostates that correspond to most populated cluster RMSD[5.0] (black profile in figure), for the whole ensemble of strings (red profile) and for the ensemble of strings corresponding to microstates having Cα-RMSD to the X-ray 1pgb structure greater than 10 Å (bleu profile). Interestingly there is not much difference between the red and bleu profiles: blue profile is a signal related to the unfolded state thus the entropy of the whole ensemble of strings mainly gives a signal resembling the unfolded state. Black curve shows very low minima that corresponding to the β-strands and the helix while the signal at the loops (GG-loops) assume comparable values with respect to those in the red/blue entropy profiles. The entropy profile corresponding to the folded cluster suggests in a further quantitative way what stated by the autopairs RMSD plot of figure 3.6 (B): the folded state for this toy protein has a very stable secondary structure (the β-sheet and the α-helix) and fluctuating loops which make the overall structure somehow "liquid". Moreover the high disorder of loops must be an intrinsic property of the sequence as it appears conserved all over the configurational space in both the folded and unfolded states. Why should that be so? In the formation of the topology of protein G one can distinguish between three main steps: firstly the formation of the central

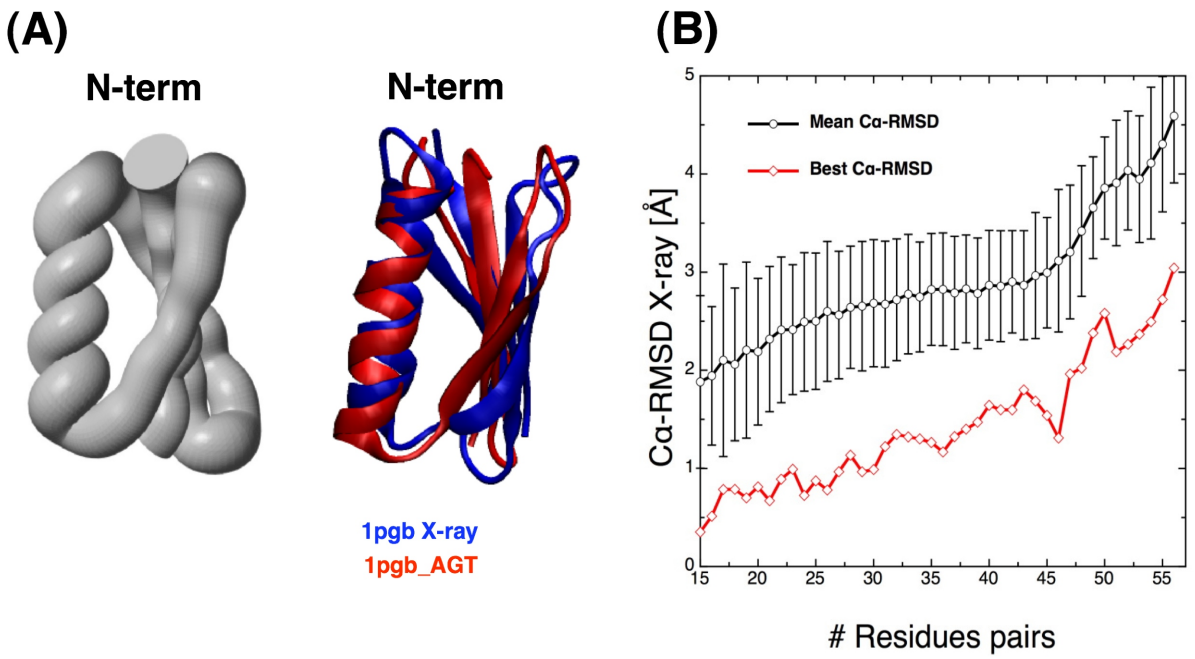

Figure 3.6: (A left) The ensemble representation of the most populated and folded cluster of structures: the mean pairwise Cα-RMSD within the cluster is 3.5 Å (all the Cα considered). (A right) The Cα-RMSD structural alignment between the folded cluster center (red ribbon) and the X-ray structure (blue ribbon) of protein G (1pgd pdb code), Cα-RMSD between the two structures is 2.6 Å where the first and last 2 Cα and the Gly residues were excluded from the calculation. (B) The Cα-RMSD with respect to the X-ray structure within the folded cluster of structures as a function of the number of Cα pairs used to compute the RMSD. Given a number of Cα pairs the structural alignment finds the best overlap between chain fragments that can also be not contiguous along the sequence. The black circles are mean Cα-RMSD values with their standard deviations while the red diamonds are the best Cα-RMSD values for a given number of Cα pairs. From the 50% up to the 85% of the Cα pairs the corresponded Cα-RMSD steadily turns out lower than 2.6 Å while the best values are around 1.5 Å. The result provides a quantitative indication of the fluctuating nature of the folded cluster and yet shows that the folded topology of protein G can be satisfied in a shallow manner, hairpins for instance can be still formed with a chain shifting up to ± 2 residues.

helix, secondly the turn/loop closures which eventually lead to the formation of the two hairpins. Helix formation is a process essentially favored by entropy as its formation can be initiated by any residue site composing the chain, in this sense the coil-helix transition can be seen as an example of non cooperative first order phase transition [Zimm and Bragg, 1958, Zimm and Bragg, 1959]. On the other hand the formation of a β-hairpin is globally disfavored by the entropy, namely it is initiated by a unique loop closure event first and promoted then by the zipping of the backbone-backbone hydrogen bonds which gradually decrease the overall enthalpy [Muñoz et al., 1997, Dinner et al., 1999, Klimov and Thirumalai, 2000]. Loop closure is a stochastic local event that can be accelerated by the disorder of the residues that

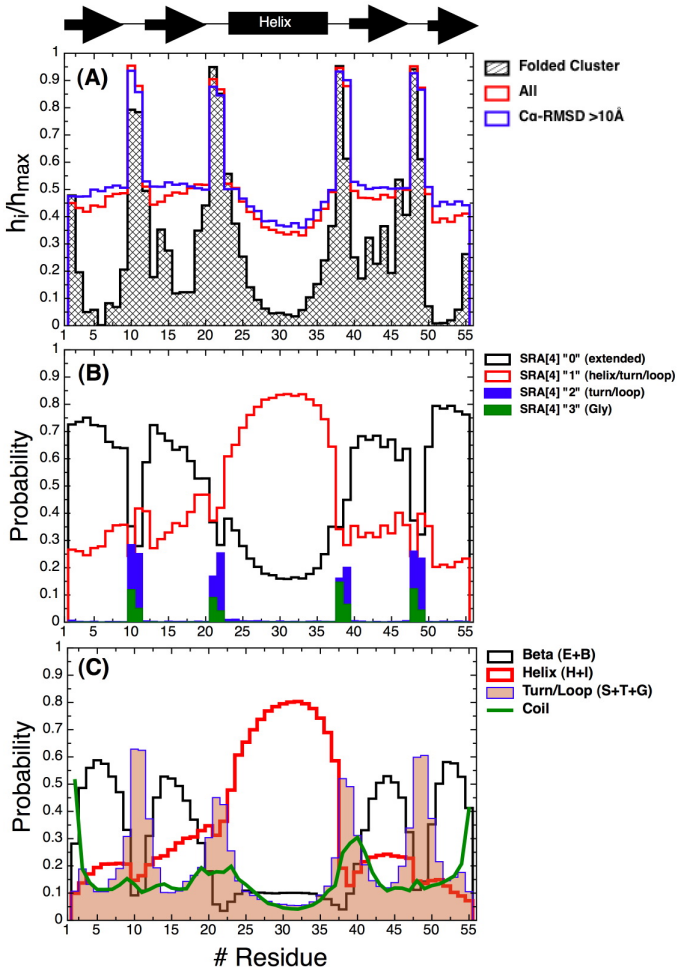

Figure 3.7: Protein 1pgb_AGT: (A) the normalized entropies per residue respectively: for the ensembles of SRA[4] string of mesostates corresponding to the RMSD[5.0] folded cluster (black curve), for the whole ensemble of strings (red curve) and for the sub-ensemble of strings such that the corresponding microstates have Cα-RMSD to the X-ray structure greater than 10 Å (bleu curve). The latter sub-ensemble of strings is clearly referred solely to the unfolded state of the protein. It is interesting to notice the slightly difference of the entropy profiles between the red and blue curve. The profile of the folded cluster is very peaked to the loop regions showing their high disorder, on the contrary the first and last strands and the helix regions appear very ordered. In (B) and (C) the statistics of SRA[4] and SSS[4] mesostates per residue are reported for the whole ensemble of strings. The SSS[4] description is based on the DSSP alphabet for secondary structure in which the following grouped states were considered: beta=E+B, helix=H+I, turn/loop=S+T+G, coil. The helix profile of SSS[4] in (C) is compatible with the helix/turn/loop profile SRA[4] in (B) meaning that the two descriptions are very similar. The peaks of disorder in (A) are based on the dihedral description, so that they essentially take into account the disorder due to the coil structures located at the N- and C terminals, and due to the different kind of turn/loop configurations that can be realized with several dihedral arrangements.

have been evolutionary selected to this purpose. Put in a different way, turn/loop closures are in principle rare events that can be accelerated through the high flexibility of specific residues (Gly-Gly, Gly-Ser, etc.)[1] involved in their formation. By means of flexibility or disorder they allow other residues to get in contact to form favorable interactions, notably backbone-backbone hydrogen bonds in a β-hairpin. An suitable metaphor to illustrate this mechanism of favorable disorder has been proposed by Caflisch in the context of the intrinsically disordered proteins: *most children (and some research scientists too) do not like to keep order in their rooms (desks), not only because it is tedious, but also because they can visually recognize and reach the toys they want to play with (papers and documents to read) more easily* [Caflisch, 2003]. As soon as a β-hairpin is formed its free energy can further decreased by the intrinsic fluctuations of entropic origin proper of the residues involved in a turn. That explains in fact figure 3.7 (A) in which the entropy peaks at the turn regions assume similar values in both unfolded and folded states. Thus the double role of the turn prone residues in both the search and the stability of the folded state might be related to the studies of Shortle about the residual native structure in the denatured state of proteins [Shortle, 1996, Shortle and Ackerman, 2001], residual native structure in the unfolded state could be originated by disordered β-turns that continuously search "their" native state. The concept that intrachain loop formation allows unfolded polypeptide chains to search for favorable interactions during folding seem in fact to be an established matter [Fierz et al., 2007, Fierz and Kiefhaber, 2007]. Turn/loop residues seem on one hand to behave like source of order in the disorder of the denatured state by facilitating the folding search, on the other hand they also act as source of disorder within the order of the folded state playing the role of stabilizers. It has been suggested in the context of studies on the origin of the genetic code that β-sheets and β-turns probably characterized "primitive" proteins and constituted the main adaptive theme promoting the origin of the genetic code [Di Giulio, 1997]. There are indications the seem to push in this direction, for instance in [Jurka and Smith, 1987b, Jurka and Smith, 1987a] it is argued that the β-turns of proteins were object for selection in the prebiotic environment and influenced the development of the genetic code and the biosynthetic pathways of amino acids , as precursor[2] amino acids are also the most abundant ones in β-turns. Yet, a study directed to clarify how the physicochemical properties of amino acids are distributed among the pairs of amino acid in precursor-product relationships found that the pairs reflect the β-sheets of proteins through the bulkiness or the "size" of amino acid [Di Giulio, 1996]. These considerations lead us to speculate that our simplified sequences and, in particular the way how they were constructed might represent templates for primitive proteins showing basic structural properties.

To further investigate the organization of the configurational spaces by means of mesoscopic descriptions such as SRA[4], the folded strings of both the proteins polyTHR_5GS and 1pgb_AGT were decomposed to perform a fragment analysis. The folded mesoscopic string for the polyTHR_5GS is the most populated one and corresponds to a six stranded β-sheet (see figure 3.3 (A)), that is `00002100000021000-00021000000210000021000000210000`. For the protein 1pgb_AGT, as a reference folded string, the one corresponding to the central structure of the most populated RMSD[5.0] cluster was taken (shown in figure 3.6 (A)), that turns out to be `10000000020000000013011111111111111100000000020000000`. Given the reference folded strings the probabilities of the possible folded contiguous substrings were estimated from the meso-

[1] Turns involving Pro residues deserve slightly different considerations due to their intrinsic rigidity which provides a strong bias the closure of β-turns [Ananthanarayanan et al., 1984, Hollośi et al., 1985, Rose et al., 1985, Wilmot and Thornton, 1988].

[2] Precursor amino acids are thought to be Asp, Glu, Ala, Gly and Ser as argued in [Wong, 1975].

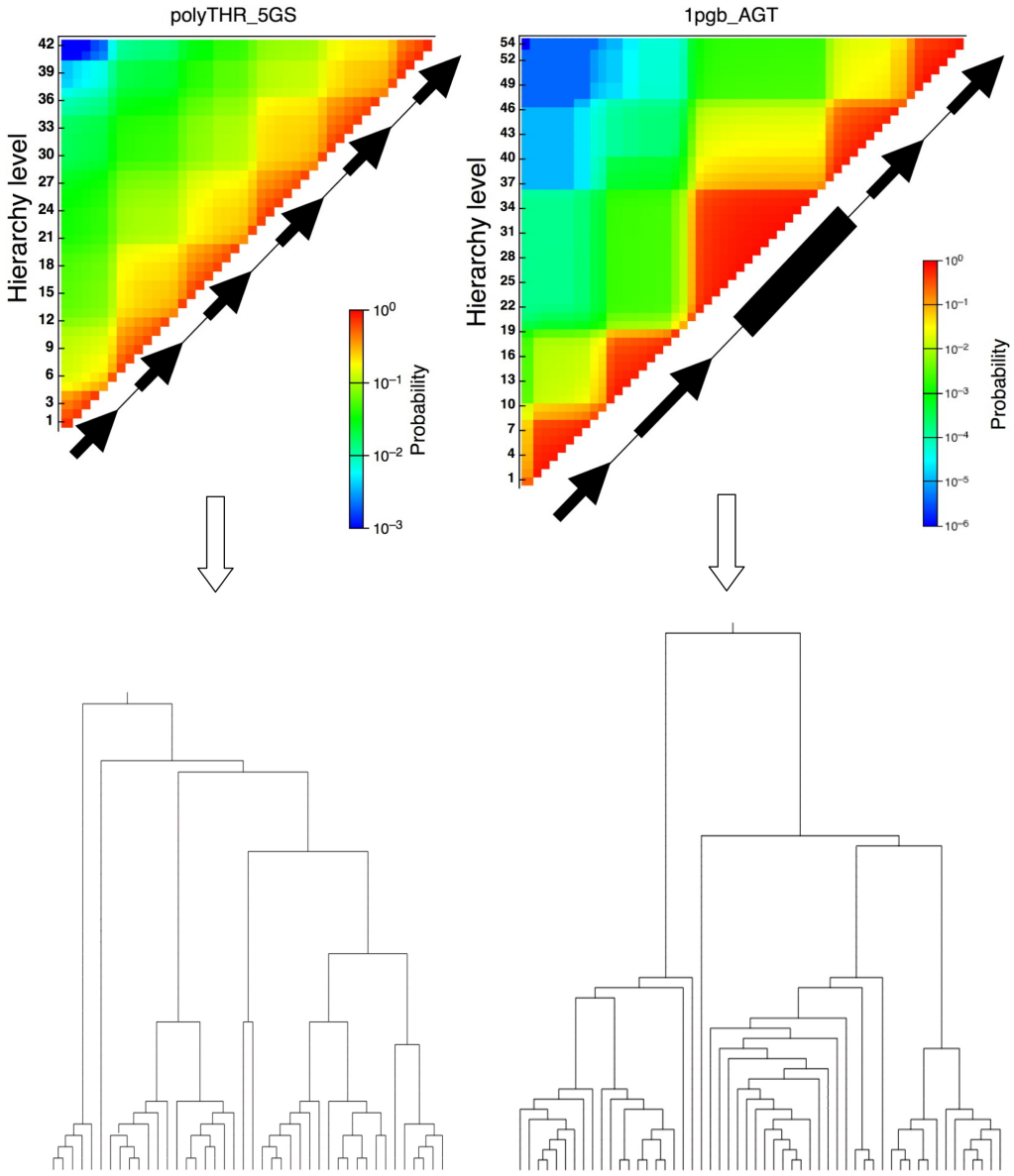

Figure 3.8: On the top the probability of the possible contiguous folded substrings is shown: an entry in the triangular map represents the estimated probability, computed on the time series, of a chain folded fragment having length going from 1 (single residue) to the full length chain. The length of a substring gives the hierarchy level on the y axes of the maps. The fragment probabilities are computed on the whole ensemble of strings discarding the full folded string, that is to reduce the bias on the substring probabilities due to the full folded string population. The maps play the role of free energy landscapes with respect to the progress variable hierarchy length, that is the number of residues that are folded. Highly ordered heterogeneity appears on the maps which can be interpreted as the existence of patterns in the ensemble of strings. From the maps hierarchy trees can be extrapolated as previously shown in chapter 2. At the lowest hierarchical level R walkers (corresponding to R residues) start a random walk, namely that 1-fragments are assembled in a certain way to gain the next status level of 2-fragments. The algorithm makes the walkers follow the maximal probability route, for instance two 1-fragments can assemble to two different 2-fragments, thus the algorithm shall choose that maximizing the 2-fragment probability. The procedure is repeated for all the hierarchies until a tree is completed by reaching the full folded string that lies on the top of the tree. The algorithm finds the maximal probability tree associated to the map.

scopic time series. Substrings are defined hierarchically from the substrings of length 1 (single residue), length 2 (contiguous residue pairs), etc. to the full string length, 42 for the polyTHR_5GS and 54 for the 1pgb_AGT. The fragment probabilities are computed on the whole ensemble of strings discarding the full folded string, that is to reduce the bias on the substring probabilities due to the full folded string population so that the probability of the full folded string is by definition 0. The results of this analysis are shown on the triangular maps of figure 3.8. Clear patterns emerges from the folded string decomposition of the proteins examined. In particular the maps represent sort of free energy landscapes with respect to the progress variable corresponding to the hierarchy length. At the lowest hierarchy level the secondary structure elements (β-strands for polyTHR_5GS and both β-strands and α-helix for 1pgb_AGT) are formed and can be viewed as fragments which freely diffuse in the configurational space. At the next hierarchical levels the secondary structure elements start to "collide" into each other by coordinating themselves thanks to the free motion of the turns Gly-Ser and Gly-Gly in a fashion resembling the diffusion-collisional model [Karplus and Weaver, 1979]. The hierarchical difference between the two proteins depends on their relative difference in secondary structure: polyTHR is essentially a repeat protein whose units are β-hairpins so that there are many ways to coordinate into each other the β-strands in order to form a one-dimensional β-sheet. Conversely 1pgb_AGT posses a α/β structure whose two hairpins have different lengths which clearly introduces an asymmetry in the folding hierarchy. However the patterns in the polyTHR_5GS map are less remarked from hierarchy level corresponding to fragments about 10 residues long, which is the length of a single βhairpin. Thus the main pattern for polyTHR_GS is simply the formation of the β-hairpins. From the maps hierarchical trees can be extracted as previously introduced in chapter 2. At the lowest hierarchical level R walkers (corresponding to R residues) start a random walk, namely that 1-fragments are assembled in a certain way to gain the next status level of 2-fragments and so on. The algorithm makes the walkers follow the maximal probability route, for example if two contiguous 1-fragments can assemble itself into two different 2-fragments (one on their left, the other on their right), thus the algorithm choose that which maximizes the 2-fragment probability, thus one has two nodes corresponding to the two initial 1-fragments which shall both link to the new node corresponding to the chosen 2-fragment. The procedure is iteratively repeated until the tree is completed by reaching the full folded string which is the top of the tree. The trees reveal a sort of syntactic structure which is parsed[3] at low level by the residues involved in turns and at higher level by super secondary assemblies. The trees are thus statistical objects which can nevertheless also interpreted as folding pathways. In this sense pattered maps may be interpreted as a lack of cooperativity in the ability of these protein to reach their folded state. In the hypothesis that the folded states here observed are molten globular (liquid-like) then the transition from an overall disordered state to a molten globular state may occur in a non-cooperative fashion [Ptitsyn and Uversky, 1994, Betz and De Grado, 1996]. On the other hand, the transition from a molten globule to a full ordered phase it is generally accepted that occurs in a slow and cooperative manner [Ptitsyn, 1995].

[3]With these terms here (syntactic and parsed) a parallel between the folded string and the syntax and punctuation of a sentence of a natural language is evoked.

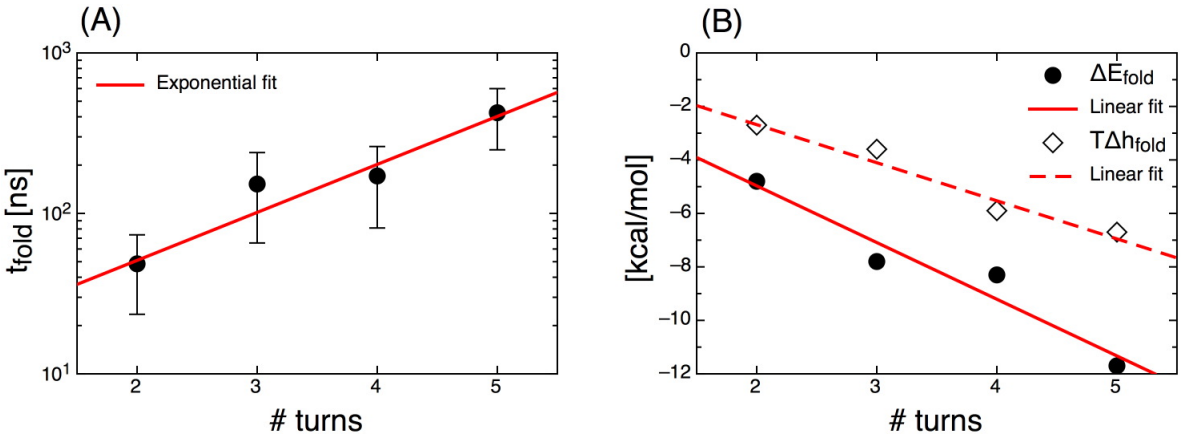

Figure 3.9: PolyTHR_xGS: in (A) the folding times are shown as a function of the number of turns of the β-stranded folded states. These times scale exponentially with the number of turns of the folded state and give e pre-exponential factor of about 12.8 ns which can be interpreted as the diffusion time of an hairpin (see text). In (B) the energy difference of the folded string (black circles) $\Delta E_{\rm fold} = \overline{E}_{\rm fold} - \overline{E}$ and the configurational entropy loss of the folded string (empty diamonds) $T\Delta h_{\rm fold} = h_{\rm fold} - h$ scaling linearly with the number of turns.

3.4.3 Folding kinetics: first passage time analysis

The proteins here studied show a strikingly fast overall kinetics which is due to the oversimplified sequences. The folded state for these sequences is kinetically very accessible as a consequence of free energy landscapes smother than those of real world proteins. Local secondary structure propensities of the amino acids dominate over the long range interactions, an aspect which is also emphasized by the hierarchical maps of figure 3.8, and consequently promote the local structural organization. It is interesting to determine whether or not these proteins posses a simple two state kinetics or more in general what kind of folding mechanism they follow. The hierarchical maps of figure 3.8 already suggest that the folding state can be reached through parallel routes although these are thermodynamic observations. The first passage time analysis FPT, as introduced in chapter 2, is a powerful method to investigate the complexity of the folding process. The FPT distributions were computed from the time series for all four proteins polyTHR_xGS where as a target state the most populated SRA[4] was considered. All four FPT distributions fit very well with a single exponential function (the distributions are not shown) so that the folding rates can be easily estimated and as well as the folding times merely taking the inverse. The single exponetiality character of the FPT distributions is a strong signature that a single overall folding free energy barrier separates the unfolded from the folded states, namely the configurational space is divided in two macro-phases. In figure 3.9 (A) the folding times of the polyTHR_xGS are shown as a function of the number of turns (or equivalently the number of β-hairpins of the folded state). These times are exponentially large with the number of turns. The data fit very well with the following simple exponential model

$$t_{\rm fold}(n) = t_0 e^{n \frac{\Delta G^{\dagger}_{\rm hairpin}}{k_B T}} \tag{3.8}$$

where the pre-exponential factor t_0 is a diffusion time of an hairpin and $\Delta G^{\dagger}_{\rm hairpin}$ is the free energy barrier to fold a single hairpin. From the data fit we obtained $t_0 \sim 12.8$ ns and $\Delta G^{\dagger}_{\rm hairpin} \sim 0.45$ kcal/mol. Essentially the model merely describe that the overall folding free energy barrier for a β-sheet composed by n hairpins is n times the folding barrier of a single hairpin. In figure 3.9 (B) the effective energy differences (black circles) $\Delta E_{\rm fold} = \overline{E}_{\rm fold} - \overline{E}$ of the folded string are reported as a function of the number of turns. Correctly, due to the fact that the energy is an extensive quantity, these energies scales linearly with with the number of turns, in particular it fits with the simple relation

$$\Delta E_{\rm fold}(n) = \Delta E_{\rm hairpin} n + {\rm const} \tag{3.9}$$

where $\Delta E_{\rm hairpin}$ turns out to be about -2.1 kcal/mol and corresponds to the enthalpy gain of a closed hairpin. Yet in figure 3.9 (B) the configurational entropy loss of the folded string is shown as a function of the number of turns. The configurational entropy loss is computed through the formula $T\Delta h_{\rm fold} = h_{\rm fold} - h$ where $h_{\rm fold}$ is the entropy of the folded string in 3.6 and h is the total Shannon entropy of the ensemble of mesoscopic strings measured in k_B units. In chapter 2 it was shown that given a certain coarse grained description, the configurational entropy loss proportional to the free energy, namely $T\Delta h_{\rm fold} \sim \Delta G_{\rm fold} = \Delta E_{\rm fold} - T\Delta S_{\rm fold}$ where $T\Delta S_{\rm fold}$ is the internal entropy difference of the folded string. Making a linear fit of the configurational entropy data of figure (B) and assuming $T\Delta h_{\rm fold} \sim \Delta G_{\rm fold}$ one has

$$\Delta G_{\rm fold}(n) = \Delta G_{\rm hairpin} n + {\rm const} \tag{3.10}$$

with $\Delta G_{\rm hairpin} \sim -1.5$ kcal/mol which is the free energy of folding of a single β-hairpin. The thermodynamic parameters here reported are compatible with those found in [Muñoz et al., 1997] in the context of β-hairpin folding.

The FPT distributions for folding and unfolding were calculated also for the 1pgb_AGT protein. As a target state for folding the most populated cluster RMSD[5.0] was used while for unfolding, a threshold of > 13 Å of the Cα-RMSD to the X-ray structure was employed to define an unfolded phase. The distributions are shown in figure 3.10. The data fit very well with a double exponential distribution where the slow phase represents folding (or unfolding) and the fast phase gives account of the diffusion in the folded state. From the fits the following parameters were obtained $t_{\rm fold} = 194 \pm 33$ ns, $t_{\rm unfold} = 64 \pm 14$ ns, $t_{\rm diff} = 6 \pm 3$ ns. Thus from the FPT data the protein seems to fit with a two state folding framework, namely there is an overall free energy barrier dividing the unfolded state from the folded. Taking from granted the two state hypothesis a $\Delta G_{\rm fold}$ can be estimated from the usual relation

$$\Delta G_{\rm fold} = k_B T \ln \frac{k_{\rm fold}}{k_{\rm unfold}} \sim -0.7 \ {\rm kcal/mol} \tag{3.11}$$

from which the population of the folded phase can be estimated through the relation

$$P_{\rm fold} = \frac{e^{\Delta G_{\rm fold}/k_B T}}{1 + e^{\Delta G_{\rm fold}/k_B T}} \sim 26\% \tag{3.12}$$

at the temperature of 330 K. The two state hypothesis does not imply that the folding reaction should proceed along a unique pathway. The existence of an overall free energy folding barrier does not mean that this barrier is unique. In other words the unfolded state can be structured and each free energy basin has its own barrier towards the folded state. The rate coming from the FPT analysis, and thus the

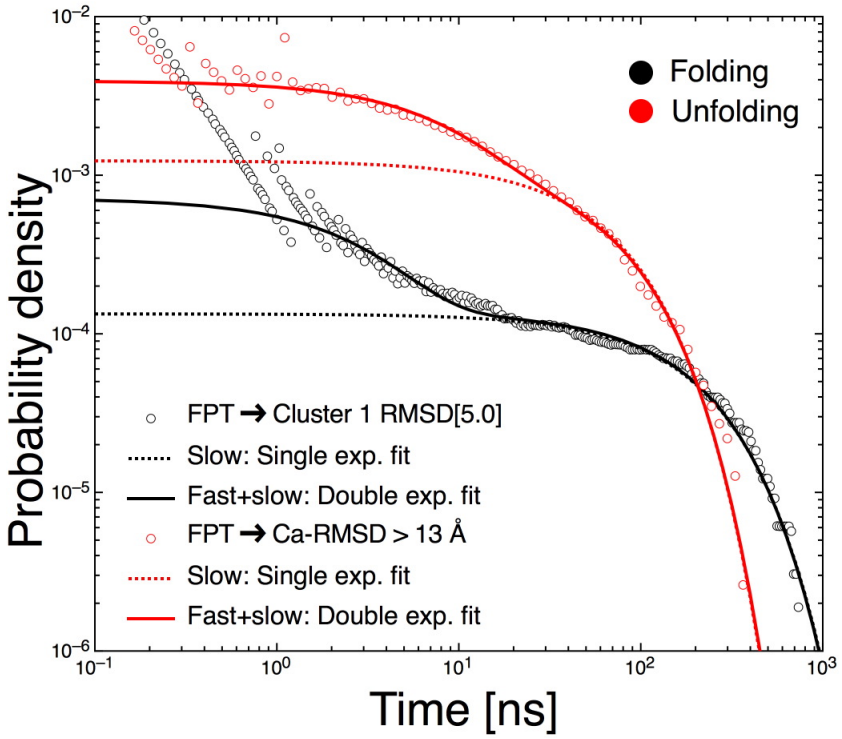

Figure 3.10: 1pgb_AGT: the FPT distributions for folding (black data) and unfolding (red data). As target state for folding the most populated cluster RMSD[5.0] (see figure 3.6 (A)) was used while a threshold of 13 Å in the time series of the Cα-RMSD to the X-ray structure was employed to define an unfolded phase. All the distributions fit very well with a double exponential function in which the slow phase corresponds to folding (unfolding) and the fast phase represents a diffusion within the folded state. From the fits it turns out $t_{\text{fold}} = 194 \pm 33$ ns, $t_{\text{unfold}} = 64 \pm 14$ ns, $t_{\text{diff}} = 6 \pm 3$ ns.

overall free energy barrier, is nothing else than the average folding barrier estimated from the ensemble of "microstates" that are not folded. Thus a two state reaction for folding does not necessarily mean an unfolded state structurally homogeneous.

3.4.4 Folding kinetics: pathways hierarchy

The analysis of the probability of all the native substrings shown in figure 3.8 can be realized also in a kinetic context. Given a reference mesoscopic string representing the folded state all possible folded substrings are considered. Following the time series of strings is thus possible to estimate the mean first passage time MFPT that is necessary to the formation of a folded substring. The formalism on how to calculate a MFPT from a time series can be found in section 2.6.2. Taking the inverse of the MFPTs of fragment formation a rate is obtained. The maps of probabilities for the native fragments already suggested that the organization of the configurational space is somehow hierarchical with the top of the hierarchy corresponding to the full folded string. Investigating the kinetics of fragment formation goes along the same direction. On the top of figure 3.11 are shown the kinetic maps of fragment formation

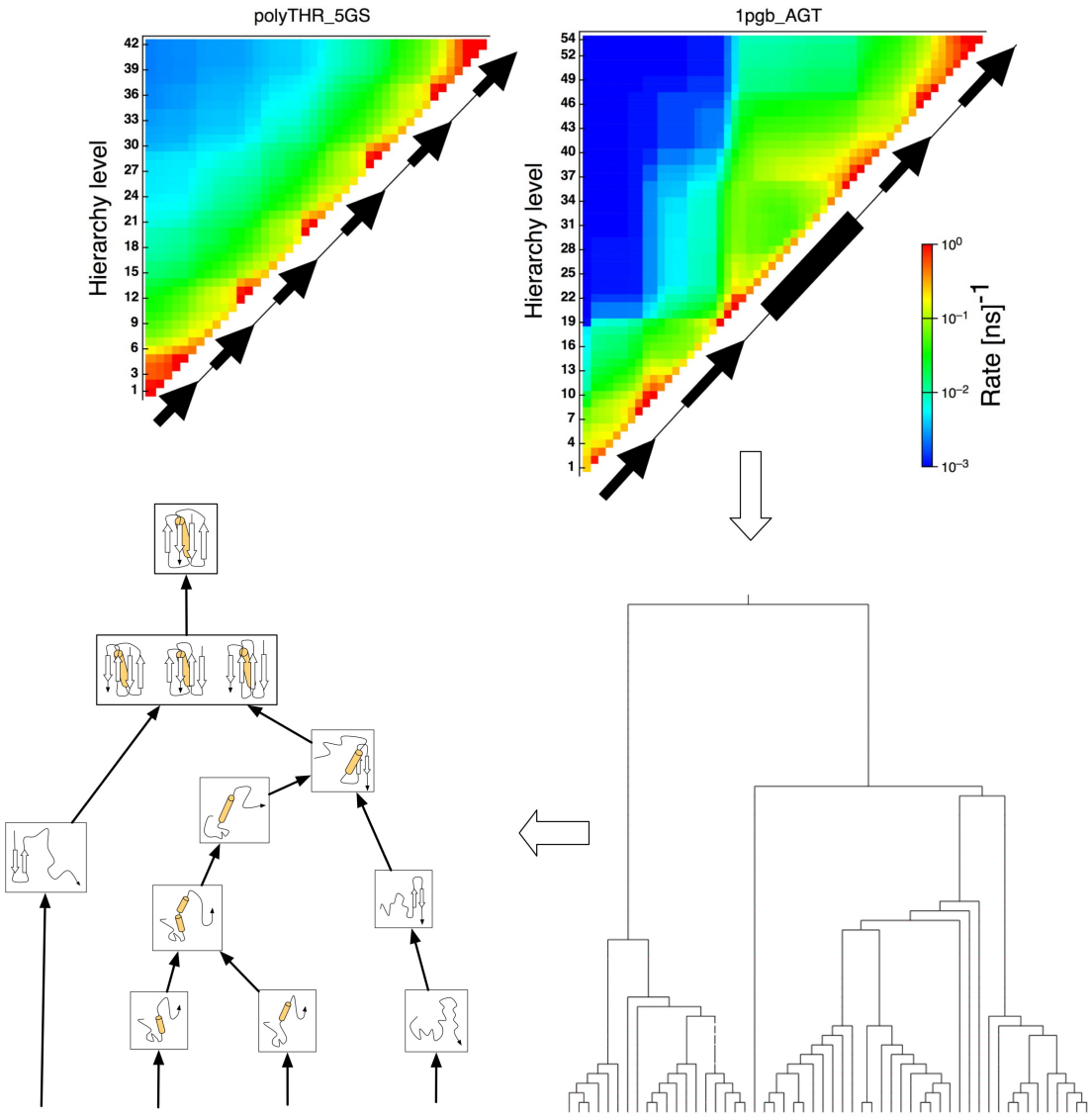

Figure 3.11: On the top the MFPTs of the all possible contiguous folded substrings is shown: an entry in the triangular map represents the estimated MFPT necessary to form a chain folded fragment having length going from 1 (single residue) to the full length chain. The length of a substring gives the hierarchy level on the y axes of the maps. The fragment MFPTs are computed on the whole ensemble of strings along the time series. The maps provide an overview of the possible folding mechanisms. The kinetic map of the polyTHR_5GS appears homogeneous suggesting that folding takes place cooperatively. Conversely the map for 1pgb_AGT looks modular with respect to the elements of secondary structure. This suggests that folding may be non-cooperative: 1st hairpin and helix plus 2nd hairpin diffuse and collide into each other. From the maps only the hierarchy trees for 1pgb_AGT could be extrapolated (bottom right in figure). At the lowest hierarchical level R walkers start a random walk, namely that 1-fragments are assembled to gain the next status level of 2-fragments. The algorithm makes the walkers follow the maximal rate route, for instance two 1-fragments can assemble into two different 2-fragments, thus the algorithm picks that which maximizes the 2-fragment formation rate. The procedure is repeated for all the hierarchy levels until a tree is completed by reaching the full folded string that lies on the top of the tree. The algorithm finds the maximal rate tree associated to the map. On bottom left an interpretation of the possible folding pathways of 1pgb_AGT.

corresponding to the polyTHR_5GS (top left in figure) and 1pgb_AGT (top right in figure). The same reference folded strings of the maps on fragment probabilities were used to construct the kinetic maps. An entry in a map represents the formation rate of a folded substring. The kinetic map of the polyTHR_5GS appears homogeneous suggesting that folding takes place cooperatively. On the other hand the map for 1pgb_AGT has a modular structure with respect the elements of secondary structure: the formation of the 1st hairpin is independent from the formation of the helix and 2nd hairpin. For both the proteins the residues involved in turns either trigger folding or play the role of mechanical joints adjusting the full formed folded state. Only for the kinetic map of the 1pgb_AGT it was possible to extrapolate a hierarchical tree which shown on the bottom right of figure 3.11. The tree provides indications on how secondary structure is hierarchical formed. Thus while for the polyTHR_5GS an all-none folding mechanism seems to prevail, for the 1pgb_AGT secondary structure formation drives the folding pathways. Possible folding pathways for the latter are schematically depicted in the bottom left of figure 3.11. The helix H is the first element of secondary structure to be formed, successively the helix and the 2nd hairpin cooperatively are shaped independently from the formation of the 1st hairpin. The assembly of 1st hairpin + (helix + 2nd hairpin) is non-cooperative, meaning that they diffuse and collide to form the full folded state. The structural topology corresponding to the folded state is formed with the 4-stranded β-sheet with both the N- and C-term internal, however other three structural topologies lead to the same native secondary structure: a) N-term external and C-term internal, b) N-term internal and C-term external, c) both N- and C-term external. Since the folded state of 1pgb_AGT posses the correct topology (that of protein G) and that it is possible to switch from a topology to another by changing two dihedral states the other three topologies might be assumed either as part of folding pathways or part of the folded basin. Notably the hierarchical tree suggest that in the former hypothesis together to the main folding pathway, in which the correct topology is shaped step by step, a parallel pathway of kind a) is present. The other two pathways b) and c) are not *a-priori* excluded.

3.4.5 Folding kinetics: Markovian dynamics and causal grouping

To further elucidate the folding mechanisms of the proteins investigated in this chapter, the time series of SRA[4] mesostates for polyTHR_xGS, and RMSD[5.0] clusters for 1pgb_AGT, were processed to construct a minimal Markov chain. The time series were processes to produce new time series of causal grouped states. The idea behind the causal grouping have been extensively introduced in section 2.7.3. Essentially the procedure is able to cure the non-Markovianity of an original time series by reassigning the states that are badly sampled to statistically meaningful states. An external parameter of the procedure is the total number of causal grouped states that at the end of the procedure one obtains. This number r_c is chosen on the basis of the analysis of the density of states distributions that are estimated from the original time series (see section 2.4.4 for the technical details). The distributions were computed on the original time series (data not shown) of SRA[4] mesostates for polyTHR_xGS and RMSD[5.0] clusters for 1pgb_AGT. The r_c number represents the number of mesostates that can be considered statistically meaningful, all the others are microscopic fluctuations, in particular the causal grouping procedure reassigns them to the statistically meaningful ones. Somehow the r_c value represents the effective number of states such that a time series can be reduced to be reproduced by a Markov chain. The r_c values, estimated from the density of mesostates of the original time series, are $\sim$310 for polyTHR_2GS, $\sim$290 for

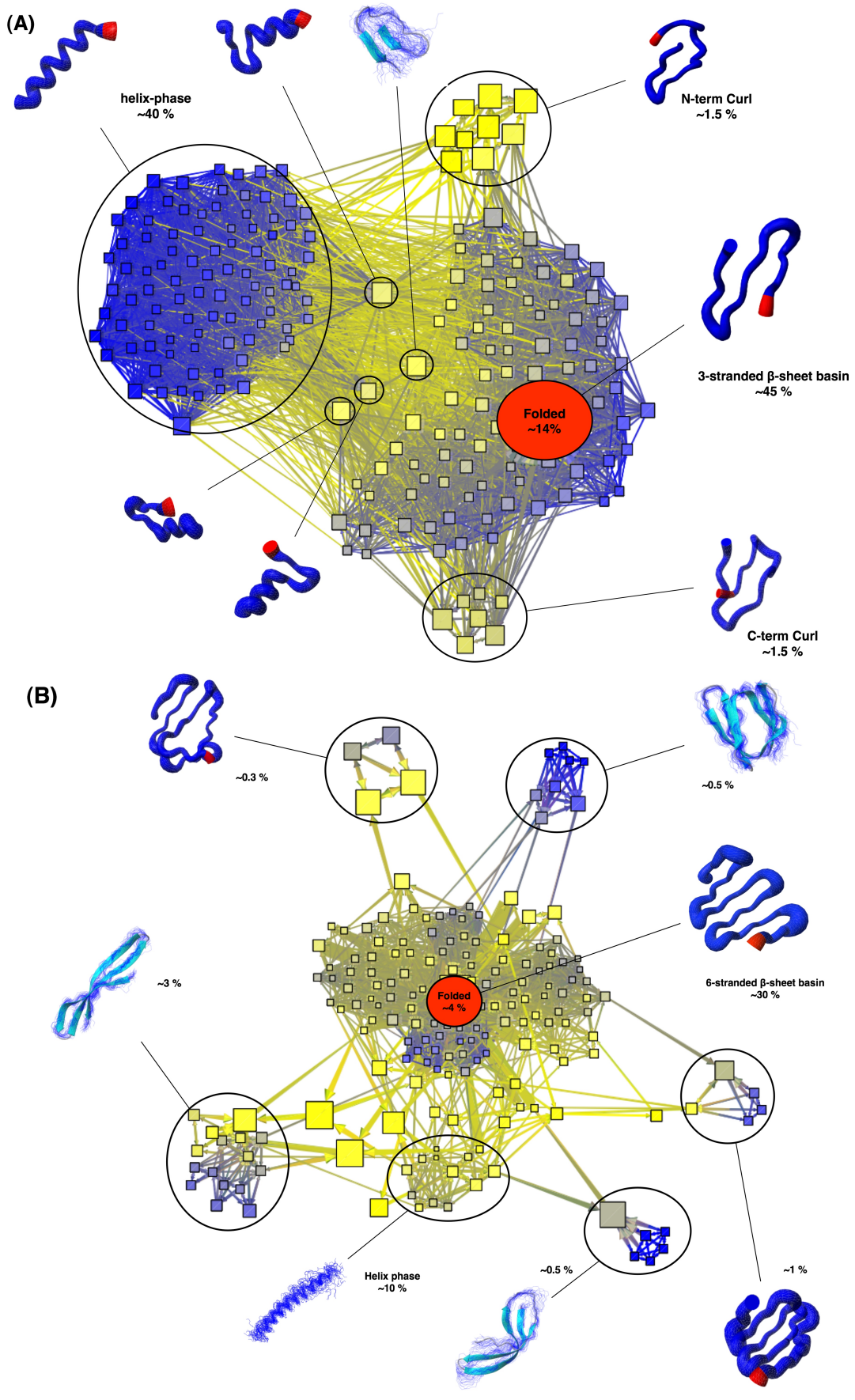

Figure 3.12: See caption of figure 3.13.

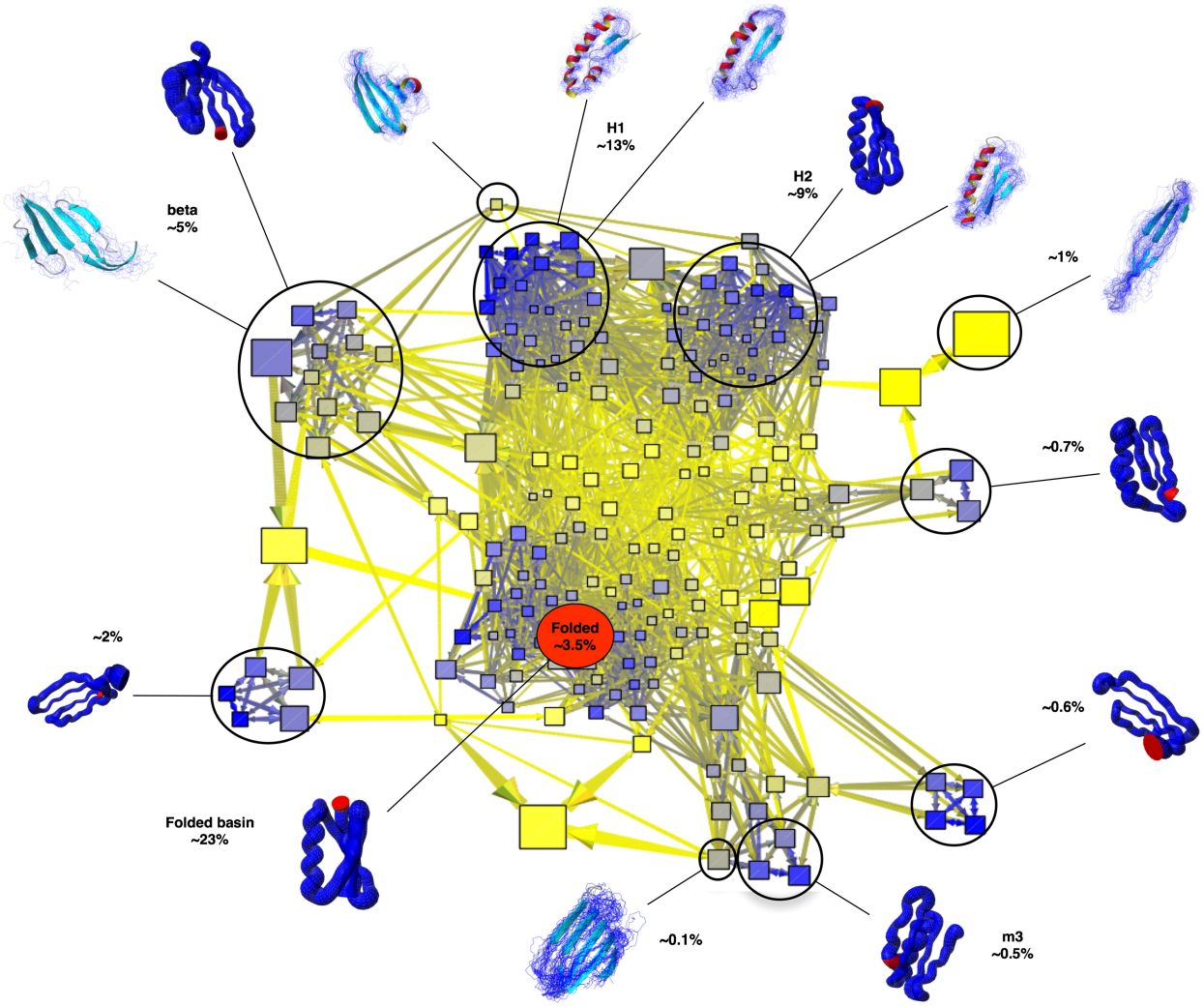

Figure 3.13: The networks corresponding to the transition matrices extrapolated from the time series of 200 causal grouped SRA[4] mesostates for the proteins polyTHR_2GS and polyTHR_5GS (figure 3.12 (A) and (B) respectively) and RMSD[5.0] mesostates for 1pgb_AGT (current figure). The figures were realized with the program Tulip [Auber, 2003] and the visualization algorithm applied is the so called spring-embedder. The links are colored according to the values of the transition matrix: darker colors correspond to high transition probabilities while clearer colors to lower values. Accordingly the color of the nodes resemble the mean value of the in-going and out-going edges. Node sizes are chosen without a numerical criteria but only to facilitate the graph reading. Cluster of nodes are grouped together into basins according to their inter-connectivities. To some nodes or basins the corresponding ensemble of structures are represented with their global populations. The graph of polyTHR_2GS is essentially divided in two main phases, a helix phase whose weight is $\sim$ 40 % and a triple-stranded β-sheet phase whose weight is $\sim$ 45 % which is the folded state. Curl-like basins are also present in both N- and C-term configuration whose weight is about 1.5 %, an aspect that is due to the symmetry of the sequence. The graph for polyTHR_5GS is much more complex due to the proliferation of non-folded β structures that also play the role of kinetic traps. The helix phase is still present though its population is reduced to about 10 %. The folded basin is large and populated about 30 %. Many "exotic" β rich basins are present. Finally the graph for 1pgb_AGT depicted in the current figure is very heterogeneous although the folded basin is clearly detectable. Two unfolded basins, H1and H2 are well defined and populated and characterized by a long helix packed respectively with a double and triple stranded β-sheet. Basin percentages indicated on networks are indicative values.

polyTHR_3GS, ∼210 for polyTHR_4GS, ∼170 for polyTHR_5GS and ∼210 for 1pgb_AGT. The reason why these numbers decrease for longer chains is that longer chains have a larger configurational space, thus the amount of statistically meaningful mesostates decrease because of a poorer statistics. A number of 200 states was used to construct the new causal grouped time series. The transition matrices were then calculated from the new causal grouped time series and used for the evolution of the Markov chains. To test the Markov hypothesis on the causal grouped time series the non-Markov flux (see section 2.7.1 for the definitions) was then calculated. The amount of non-Markov flux was found to be below the 1 % for all the causal grouped time series, which means that the Markov approximation can be safely adopted with an error below 1 %[4]. The network visualization of the transition matrix is particularly useful to investigate the kinetics of a molecule in its configurational space. When associated with a Markovian dynamics it represents the ensemble of possible mesoscopic pathways that a protein can make. The word mesoscopic is both referred to the adopted description of the configurational space and to the time step at which each transition is considered. It is important to distinguish between mesoscopic and macroscopic time scales. The latter are consequence of the equilibrium behavior of the Markov chain that is constructed on the transition matrix. In figures 3.12 and 3.13 the networks corresponding to the transition matrices of polyTHR_2GS, polyTHR_5GS and 1pgb_AGT are shown respectively. The program Tulip was used to visualize the networks [Auber, 2003] and the visualization algorithm is the so called spring-embedder. The links are colored according to the values of the transition matrix: darker colors correspond to high transition probabilities while clearer colors to lower values. Accordingly the color of the nodes resemble the mean value of the in-going and out-going edges. Node sizes are chosen without a numerical criteria but only to facilitate the graph reading. The network of the polyTHR_2GS (figure 3.12 (A)) appears quite simple. Two competitive tight phases are represented, an unfolded helix phase and a folded three stranded β-sheet. The population of these two phase are approximatively 40 % the former and 45 % the latter. As for the GSGS peptide two "side" basins such as the curl-like N-term and C-term, play the role of kinetic traps. Because of the symmetry in the peptide sequence these two basins have similar population, about 1.5 %. The network is smooth and all the basins are well grouped and divided. That means that the real free energy landscape of the molecule is as well as simple and smooth. The extreme simplicity of the sequence and the amphiphilicity of the THR residues make the configurational space modeled by turns and secondary structure propensities. In this sense the system is perfectly two state: the unfolded state is the helix phase which is stabilized mainly by entropy while the folded state in the β which is stabilized by secondary interactions. Causal grouped mesostates in between the two phases are paradigmatic of the folding mechanism: helices are broken by both turns and disrupted by the formation of the β-hairpin contacts. The network corresponding to the polyTHR_5GS (figure 3.12 (B)) shows a higher complexity. The 6-stranded β-sheet causal grouped mesostate (4 %) is the center of a wide basin that approximatively weights the 30 %. This basin is the folded basin and it is stabilized both by the enthalpy (secondary interactions) and entropy (the fluctuations of all the hairpins composing the 6-strand). The folded basin is surrounded by many other basins. The helical basin is made by metastable configurations slowly communicating between them (links are colored in yellow, meaning low values in the transition matrix). To the helical basin contribute many disordered helices and globally it weights about the 10 %. All the rest of the basins are essentially kinetic traps, mainly

[4] In average only less than 1 % of the double step transition which are predicted from the Markov approximation were not directly observed in the original causal grouped time series.

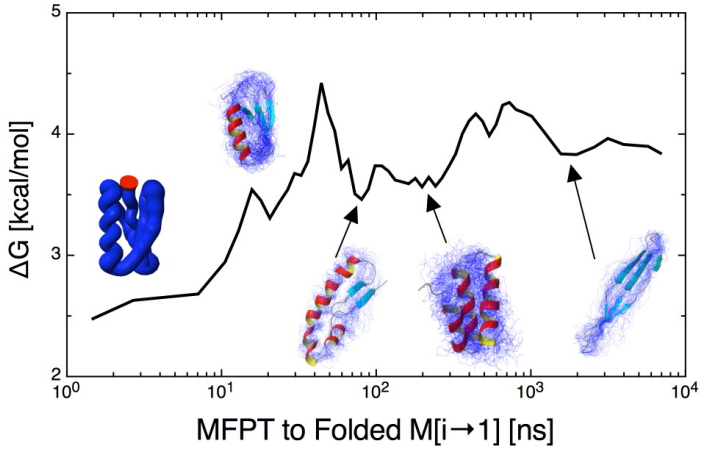

Figure 3.14: A uni-dimesional free energy profile for the protein 1pgb_AGT from the causal grouped mesostates RMSD [5.0]. The reaction coordinate is the calculated equilibrium MFPT from any mesostate to the folded through the evolution of the Markov chain on the causal mesostates. They are the values $M_{j\to folded}$ of the extrapolated matrix of the MFPTs. The values in the y axes are stability ΔG values extrapolated from the main diagonal of the MFPT matrix. The values are reported for less than an additive constant. Two main minima separated by a barrier are evident. The main unfolded basin is separated from the folded though about 1 kcal/mol barrier. A far basin relaxing very slowly to the folded state represents fibrous states.

fiber-like or compact β rich configurations, which are disfavored by the entropy. All these β configurations essentially satisfy all possible configurations that a polypeptide having 5 turns could realize.

Finally the network of 1pgb_AGT (figure 3.13) is also remarkable in its complexity. Causal states are constructed from the time series of the RMSD[5.0] clusters. The heterogeneity of the mesoscopic pathways is here also evident. In particular the folded basin is clearly separated from the rest of causal states. The main causal folded node weights 3.3 % while the folded basin approximatively weights about 23 %. The latter percentage is a roughly estimated summing up all the weights of the mesostates that relax to the folded state in less then 10 ns (data not shown). By looking into the structural details of the folded basin one can see that, while the very bottom of the basin posses the structural topology of protein G, the basin border have the correct secondary structure but some topological "imperfections", notably the 1st hairpin have the N-term in the outer of the 4-stranded β-sheet. The border of the folded basin is in a fast equilibrium with its bottom. The main basins of the unfolded state are those indicated in figure with H1 and H2, characterized by a long helix packed with either with 2- or 3-stranded β-sheet. These basins are stabilized mainly by the conformational entropy and are enthalpy disfavored with respect the folded basin of about 2 kcal/mol. Other basins play the role of kinetic traps or misfolded states, full β configurations in which the helix is almost never formed or fiber-like non equilibrium states. The basin indicated with m3 in figure is an example of state having the correct secondary structure of protein G but having both the hairpins forming the 4-stranded β-sheet anti-parallel/parallel/anti-parallel (the

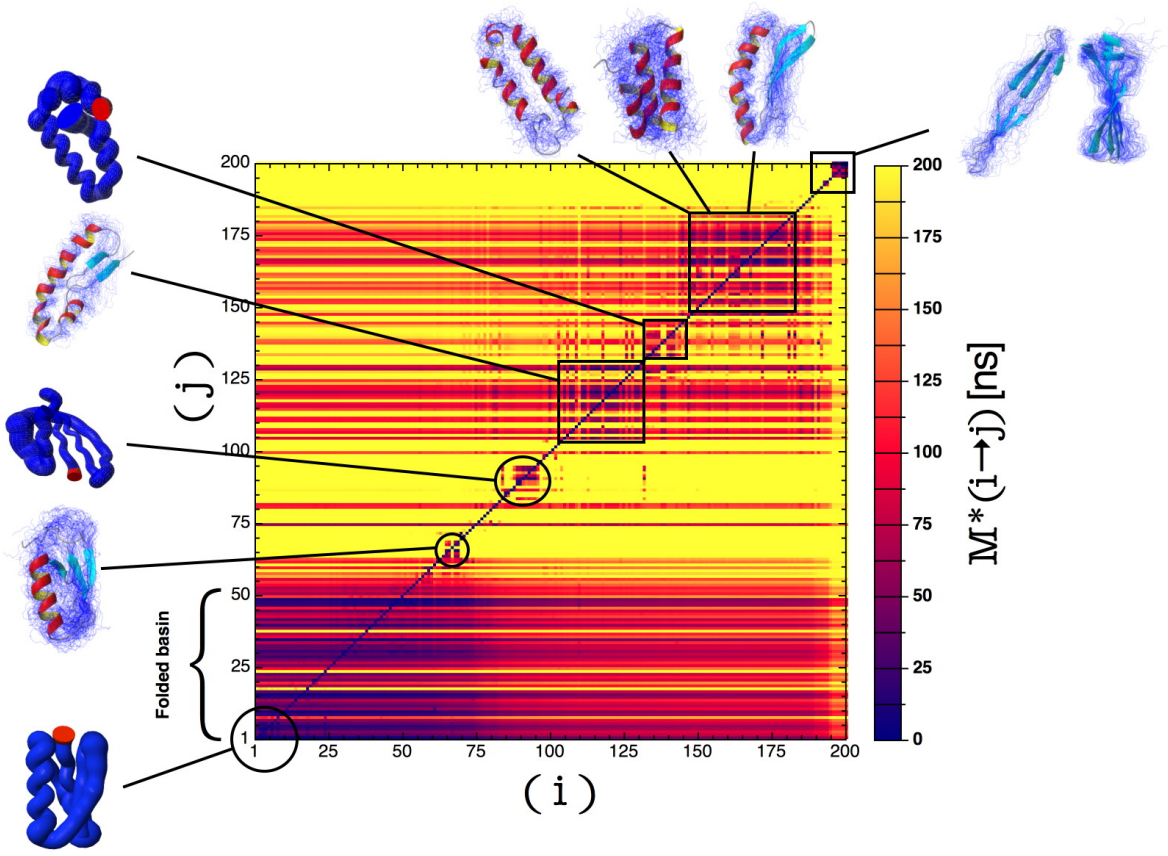

Figure 3.15: The reordered MFPT matrix $M^*_{i\to j}$ for the protein 1pgb_AGT based on causal grouped mesostates from RMSD[5.0]. An entry on the matrix gives the MFPT for the equilibrium transition $i \to j$. Horizontal bands are equilibrium transitions from all the is to a specific j. The index (i,j) are ordered from 1 fastest relaxation to the folded state to 200 slowest relaxation to the folded state. The folded basin is composed by the dense bands going from 1 to about 50. The fact that the bands are dense means the folded state can be reached from many gateways in about 150/200 ns, in particular the Markovian MFPTs analysis confirms that an overall folding free energy barrier separates the unfolded state from folded. Most kinetically far states are the fiber-like states that are mainly non-equilibrium states (the system visits them barely once and never returns there). The proper unfolded state is thus populated by helices in several combinations (helix bundles, α/β, etc.) rapidly exchanging between them.

folded configuration is parallel/parallel/parallel). This basin is far from the folded state so that it can be assumed as an example of misfolded state.

How do the basins interchange between them on equilibrium time scales? In other words, is folding the result of a multi-basin diffusion in the unfolded state or the folded state is attained directly. The FPT analysis has shown that an overall free energy barrier separates the unfolded phase from the folded. Does that mean that the basins in the unfolded state can inter-convert? The evolution of the Markov

chain constructed on the causal grouped states can answer to these question. In section 2.7.6 it was shown how to calculate the matrix M_{ij} of all the mean first passage times MFPT from the transition matrix. An entry in the matrix M_{ij} represents the mean time that is necessary to reach the state j starting from i on an equilibrium time scale. We calculated the matrix M_{ij} for the Markov chain corresponding to the causal states of the protein 1pgb_AGT with 200 nodes. As usual the causal state with id 1 correspond to the folded state (namely the most populated state) therefore the first row of the matrix $M_{i\to 1}$ provides the MFPTs from any starting causal state to the folded state. As it was shown in section 2.7.6 we consider the MFPT to the folded state as a sort of reaction coordinate. In figure 3.14 we show a uni-dimensional free energy profile in which the reaction coordinate is the MFPT to the folded state. This coordinate corresponds to the first row of the matrix M of the MFPTs, namely $M_{j\to 1}$ for all the js. The values in the y axes of figure 3.14 are ΔG stability values of the causal grouped mesostates, extrapolated from the main diagonal of the MFPT matrix. The main diagonal of the M matrix gives the mean revisiting times $M_{i\to i}$ of the mesostates. These times satisfy to the the relation $M_{i\to i} \sim t_0\, e^{\Delta G_i/k_BT}$ where t_0 is a reconfiguration time scale and ΔG_i is the mean free energy difference between the mesostate i and all the others. The revisiting times are equilibrium extrapolations thus, they are direct consequence of the overall stability of the mesostates. Interestingly the about 1 kcal/mol barrier from the main unfolded state to the folded state is consistent with that found from the FPT calculations.

To investigate the equilibrium connectivity of the causal mesostates through the analysis of the M_{ij} matrix we reorder its indexes in such a way that the low indexes (from 1) are those mesostates possessing low MFPT to the folded state while large indexes have larger values of MFPT to the folded state. We call the reordered matrix M^* so that the first row satisfies to the inequalities $M^*_{1\to 1} \leqslant M^*_{2\to 1} \leqslant \cdots \leqslant M^*_{200\to 1}$. The matrix is shown in figure 3.15 and its band structure provide useful informations on the folding mechanisms. First: the folded state is layered from index 1 to about 50 which essentially means that there are many gateways to access the folded state and that there is an overall folding free energy barrier separating what is unfolded from what folded is. That confirms what the FPT analysis already suggested, namely that macroscopically this protein is a two state folder with an exponential folding time distribution. Thus no matter where the system is unfolded the mean time to fold is always about 200 ns. Second: a highly fluctuating folded basin makes it easily accessible. A useful comparison is the same MFPT matrix that was obtained for the GSGS peptide. There the folded state is much less kinetically layered and more stable. Interestingly the unfolded state is helix rich. Many combinations of helices structures (helix bundles, α/β, etc.) are rapidly exchanging between them, that is the further band structure of the MFPT matrix that runs from index 100 to 175. These states correspond to the basins indicated with H1 and H2 in the network of figure 3.13. Likely the presence of a band structure is a signature that these two basins are part of bigger macro-basin. The slowest states relaxing to the folded state are those fiber-like which are essentially non equilibrium states, that is the system visits them once and never returns there. Thus the picture arising from the current analysis is that although the folding kinetics can macroscopically be regarded as two state, a folded basin whose stability is about a k_BT increases the number of parallel routes towards the folded states. Thus a lack of stability produces more productive folding pathways but at the same time increases the diffusion among unfolded basins.

3.5 Conclusions

What can one learn from toy model proteins? The simplified scheme with which the sequences were constructed has on one hand reduced the intrinsic frustration of the free energy landscape and made more fluid-like the dynamics in the configurational space. All the barriers are overall reduced in comparison to the natural proteins. This has the consequence that first, the local interactions play a leading role in shaping the configurational space, and second, the folded state posses a higher kinetic accessibility; third, many other basins are accessible too but folding is still under kinetic control. The system diffuses in a configurational space in which folding appears not under a thermodynamic control, namely that the folded state can be easily reached but its stability is not granted. Many of the β rich basins that were observed for the 1pgb_AGT protein might be potentially dangerous states as they could promote the β-amyloid aggregation. Interestingly, coarse grained simulations of (poly)peptide aggregation indicate that a minor increase (< 1 kcal/mol) in relative stability of a β-aggregation prone state, can result in a dramatic acceleration of fibril formation rates [Pellarin et al., 2007].

Experimentally, it has been shown in [Ramirez-Alvarado et al., 2000] that by reducing its stability and under certain experimental conditions, some variants of B1 Ig-binding domain of protein G form fibrils with high reproducibility. By controlling the stability of the protein, mutations or variation of the experimental conditions, it was possible to modulate the ability of the protein to form fibrils. For all of the variants, they found that the key requirement for fibril formation was to choose conditions in which the population of intermediate states present during the unfolding transition was maximized. Notably they also suggested in [Ramirez-Alvarado and Regan, 2002] that the overall protein stability is the key determinant for amyloid formation and not the specific location of destabilizing mutations. Consequently on the basis of the results here presented we suggest that the evolution of protein sequences has been directed towards a double purpose: the optimization of protein function (and stability) on one hand and the elimination of dangerous intermediate states that would compete with the folded state. Reduced alphabets of amino acids can be suitable to define elementary folds but they do not encode the sufficient complexity such that these optimization prescriptions could be evolutionary achieved.

The configurational space is modeled mainly only by local interactions, notably interactions that are responsible of the secondary structure patterns. Moreover in our toy proteins no explicit hydrophobic core is present so that the attained folds are characterized by high flexibility and lack of specific tertiary long range contacts resembling a molten globular structure. It is also interesting to notice how in our simplified proteins the disorder of the residues responsible for making turn or loop, play a double role: first they facilitate the folding search acting as mechanical joints to allows the native interactions to be formed and second they work as entropic stabilizers of the folded state. The explored configurational space appear modeled by all structural patterns that are compatible with the secondary structure that a polypeptide 56 residue long can assume. To the broad folded basin not only contribute configurations with the correct folded topology but also states which posses a folded-like secondary structure pattern. These states fast inter-convert to the topologically correct folded state by paying a small entropic price. In the main unfolded state several helical states, including helix bundle, share the same macro-basin (see figure 3.14) which is separated by an overall 1 kcal/mol free energy barrier from the folded state.

This study provide indications that low complexity amino acid alphabets might be already able to encode complex protein topologies such as α/β proteins. If further confirms will come from the ex-

perimental side that would lead to the conclusion that high complexity alphabets are the product of an evolutionary path biased more in favor of protein function than in protein structure. This conclusion has been also proposed in the context of designed proteins, such as the Top7 α/β that shows a strong non-cooperative folding [Watters et al., 2007]. Even in the early works of Finkelstein and Pitsin when the folding patterns were investigated it was suggested that the limited diversity of folding patterns might be the result of physical limitation rather than an evolutionary divergence or functional convergence of proteins [Ptitsyn and Finkelstein, 1980, Finkelstein and Ptitsyn, 1987].

Finally, simplified proteins can be a powerful tool to computationally investigate the folding mechanisms of small size proteins (60 residues) by means of molecular dynamics simulations. The fact that spontaneous folding was observed in our simulations is also a remarkable aspect of this study. To our record there is no similar study that have showed spontaneous folding in silico for an all-atom model except for Go models that are native-centric. The simplifying strategy here employed resulted in a overall decrease of the energetic frustration of the system which has led to a smoothed free energy landscape. Consequently the overall folding rate of the system is made accessible for computational investigations. Although the folding mechanisms of the simplified proteins might not be directly linked to those of natural proteins, we nevertheless believe that molecular dynamics simulations of reversible folding for small size proteins can be an interesting tool for testing folding paradigms.

4 How does a simplified-sequence protein fold?

(Submitted manuscript)

How does a simplified-sequence protein fold?

Enrico Guarnera, Riccardo Pellarin, and Amedeo Caflisch*

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
FAX 0041 44 635 68 62

(Dated: February 9, 2009)

We have simplified the sequence of a 56-residue α/β fold (the immunoglobulin-binding domain of protein G) by replacing it with polyalanine, polythreonine, and diglycine segments at regions of the sequence that in the folded structure are α-helical, β-strand, and turns, respectively. Remarkably, multiple folding and unfolding events are observed in a 15-μs molecular dynamics simulation at 330 K. The most stable state (populated at about 20%) of the simplified-sequence variant of protein G has the same α/β topology as the wild type but shows the characteristics of a molten globule, i.e., loose contacts among side chains and lack of a specific hydrophobic core. The unfolded state is heterogeneous and includes a variety of α/β topologies but also fully α-helical and fully β-sheet structures. Transitions within the denatured state are very fast, and the molten-globule state is reached in less than 1 μs by a framework mechanism of folding with multiple pathways. The simulation results suggest that evolution has enriched the primordial, low-complexity alphabet of amino acids not only for optimizing protein function (i.e., for stabilizing a folded state with specific tertiary interactions) but also to prevent the accumulation of misfolded states rich in β-sheet structures that are prone to aggregate.

Keywords: Markov approximation, causal grouping, reduced amino acid alphabet, evolution, folding pathways, molten globule, implicit solvent, molecular dynamics simulations

Author Summary

Ancient proteins consisted of a small subset of the 20 types of amino acids present in today's proteins. We have simplified the sequence of a modern protein (having an α/β topology) by employing only three types of residues to make it similar to a prebiotic protein. Using atomistic molecular dynamics simulations we have shown that it folds rapidly and reversibly to the structure of the modern protein. The simulations results indicate that the denatured state of the putatively primordial protein is very heterogeneous as it includes a variety of α/β conformations but also fully α-helical and fully β-sheet structures. Moreover, its folded state is fluid, i.e., resembles a molten globule. We conclude that evolution has enriched a primordial amino acid alphabet not only to optimize protein function but also to reduce the presence of misfolded conformations potentially prone to induce the formation of non-functional or even pathological aggregates.

I. INTRODUCTION

Proteins fold by a complex transition from a very broad ensemble of unfolded conformations to the well-defined native state, which is the functional structure. The complexity originates from the many degrees of freedom and the delicate balance of enthalpic and entropic contributions to the free energy from the polypeptide chain and solvent molecules (1–3). Thus, despite folding involves one single polypeptide (in aqueous solvent) the folding process is described more appropriately as a phase transition rather than a simple chemical reaction (3; 4).

Evolution has selected sequences for specific biological functions, which, except for the na-

*corresponding author: caflisch@bioc.uzh.ch

tively unfolded proteins, require a thermodynamically stable folded structure (5). Although folding efficiency is not under direct evolutionary pressure, fast folding (i.e., in the microsecond to second time scale) is necessary for many biological functions that have to be fine-tuned in time, such as signal transduction and rapid adaptation to changes in the environment. Concerning a stable functional state, it has been suggested that a sufficiently high diversity of interactions is required for folding to a unique state with an energy much more favorable than "decoy" structures (6; 7). Diversity of interactions requires an heterogeneous amino acid alphabet. Theoretical analysis and computer simulations have suggested that selection of sequences that yield a native conformation with a pronounced energy minimum, i.e., an energy gap with respect to other structures, solves the problem of kinetic accessibility of the native conformation (8). Furthermore, by a comprehensive computational analysis of the folding cooperativity in several widely used lattice models, it was observed that the model based on a 20-letter alphabet is the most cooperative while 2- and 3-letter models are much less cooperative (9).

On the experimental side, random libraries of sequences with only three types of amino acids (leucine, glutamine, and arginine) have been expressed in *E. coli* (10–12). By means of circular dichroism measurements, only 1% of the sequences were shown to fold. These results led the authors to conclude that the key elements of protein design is the proper placement of hydrophobic residues along the polypeptide chain to ensure the formation of a well packed hydrophobic core. In another experimental study the sequence of the SH3 domain was simplified by using only five types of amino acids (glycine, alanine, isoleucine, lysine and glutamate) (13). The study was conducted using the phage-display technique to select for native function. Despite the dramatic change in sequence, the folding rates of the simplified versions of the SH3 protein were very close to the folding rate of the wild type. Moreover, NMR analysis provided evidence of a well packed core consistent with the thermodynamic stability of the folded state.

It is still very far from routine to simulate reversible folding of (even small) proteins by transferable potentials because of the time scales involved (microseconds to seconds) as well as the systematic error of the atomistic model. Here, we attack the complexity of the folding process by designing and simulating a putatively primordial protein, a variant of the immunoglobulin-binding domain of protein G with a simplified sequence (termed protein ssG hereafter). The simplified (i.e., low complexity) sequence of protein ssG consists of only three types of residues, glycine, alanine and threonine, which are distributed to preserve the secondary structure propensity of the wild-type sequence. The present study was inspired by the following questions: What is the folding mechanism of a protein with simplified sequence? Is its folded state topologically equivalent to the one of the wild type and is it uniquely defined? Is its denatured state heterogeneous, i.e., does it contain native and/or non-native secondary structure elements and topologies? Are there misfolded states that might promote aggregation? The simulation results indicate that the protein ssG folds rapidly and reversibly to the native topology of the wild type but has a fluid-like folded state devoid of specific hydrophobic contacts. Furthermore, the strong propensity for regular secondary structure formation results in a framework model of folding with parallel pathways. Notably, the heterogeneous unfolded state ensemble of protein ssG includes fully β-sheet traps, which are likely to be aggregation-prone.

II. METHODS

A. Reduced amino acid alphabet and simplified sequence of protein G

A necessary condition for protein-like sequences, namely sequences resulting in an energy gap between folded state and "decoys", is that the effective number of amino acid types m_{eff} is larger than the number of conformations per residue γ (6). Assuming that a single residue can be found in three states of secondary structure, helix, beta and turn/loop, we hypothesized that the condition $m_{\text{eff}} > \gamma$ might hold for native topologies mainly defined by secondary contacts, adopting an extremely simpli-

fied alphabet of solely three amino acids. In other words, our Ansatz is that it is sufficient to choose three amino acids specifically prone to form the aforementioned secondary structure to reproduce the starting fold. Thus, to enforce secondary structure propensity and remove frustration the sequence of protein G was simplified into only alanines, threonines, and glycines at segments that in the folded structure are α-helical (residues 23-37), β-strand (residues 1-9, 12-20, 40-47 and 50-56), and turns, respectively. Threonine was chosen not only because it is a moderately β-prone residue but also to counterbalance the hydrophobicity of alanine and glycine. Moreover, threonine is the most abundant residue in the wild-type sequence and it is present in 24% of β-strand segments. Table I shows the sequences of wild-type protein G and the variant protein ssG. The sequence identity is only 23% and the 13 identical residues are almost uniformly distributed along the 56-residue sequence except for Thr_{16}-Thr_{17}-Thr_{18} in the second strand of the N-terminal β-hairpin.

B. Molecular dynamics simulations

All simulations and most of the analysis of the trajectories were performed with the program CHARMM (14); the rest of the analysis was done with the program WORDOM (15), which is particularly efficient in handling large sets of trajectories. Protein ssG was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field (16) with the default cutoff of 7.5 Å for the nonbonding interactions). A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent (17). The weakest approximation in SAS is the neutralization of charged groups but these groups are not present in the ssG variant of protein G which has one polyAla and two polyThr-Gly-Gly-polyThr segments spanning the central α-helix and terminal β-hairpins, respectively. In other words, if one neglects the two termini there are only four different functional groups in protein ssG: Secondary amide and methylene in backbone, and methyl and hydroxyl in the side chains.

Despite the neglect of collisions with water molecules (frictional effects) in the simulations with the implicit solvent model, the *relative* rates of folding for different secondary structural elements are comparable with the values observed experimentally; i.e., helices fold in about 1 ns (18), β-hairpins in about 10 ns (18) and triple-stranded β-sheets in about 100 ns (19), while the experimental values are $\sim$0.1 μs, $\sim$1 μs and $\sim$10 μs , respectively (20; 21). A 15-μs molecular dynamics simulation of protein ssG was performed at 330 K which is a temperature at which the unfolded and molten-globule state are significantly populated (see Results). The temperature was kept constant by means of the Berendsen thermostat with time constant of 5 ps. A time step of 2 fs was used and the coordinates were saved every 20 ps for a total of 750000 snapshots. The simulation required about 18 weeks of a 2.8 Ghz Athlon CPU.

C. Coarse-graining of conformational space

A molecular dynamics trajectory is a long series of microscopic configurations each of which is visited only once. For this reason the analysis of the system needs a preliminary coarse-graining of the trajectory that allows the grouping of similar structures. There are several meaningful approaches that are known to efficiently achieve coarse-graining. For a system like protein ssG, root mean square deviation (RMSD) clustering and secondary structural symbolization are reasonable choices (22–25). In this work both approaches were used for different type of analysis. For the C_α-RMSD clustering we adopted the quality-threshold algorithm (26) in the version implemented in the program WORDOM (15). The choice of a clustering cutoff of 5 Å was particularly effective in capturing the salient structural motifs of the stable states of the ssG protein due to the high flexibility of the main chain. The number of clusters with two or more snapshots are 3124 and include 77% of the total sampling while the remaining 23% are unassigned structures. Not all the clusters can be considered statistically significant due to the finite size of the sampling. The statistical significance can be evaluated from the distribution of cluster sizes, namely the number of clusters $k(n)$ with n members (the cluster size distribution is shown in

Suppl. Mat. Fig. 1). The profile of this distribution follows a lognormal dependence, whose mean value gives the order of magnitude of the cluster sizes cutoff above which the clusters are statistically significant, which is $\gtrsim 100$.

As an alternative to RMSD clustering, strings of secondary structure can be employed to "symbolize" the trajectory of the protein ssG. According to the DSSP (27) code each residue can have one of eight symbols `-` (coil), `E` (extended strand in a β ladder), `S` (bend), `T` (hydrogen bonded turn), `B` (residue in isolated β-bridge), `G` (3_{10} helix), `H` (α helix), `I` (π helix). Hence each protein conformation is identified by an octal string (string of secondary structure, SSS[8]). The maximum number of strings for a polypeptide chain of 56 amino acid is $8^{54} \sim 10^{48}$, as the N-terminal and C-terminal residues have no assigned secondary structure.

D. Markov chain approach, causal grouping and mean first passage times (MFPT)

From the time series of C_α-RMSD clusters a one-step transition matrix $\mathbf{T}(\tau)$ of conditional probabilities can be estimated by using the relation

$$T_{ij}(\tau) = P_{ij}^{eq}(\tau)/P_i^{eq} \simeq n_{ij}(\tau)/n_i \qquad (1)$$

where the indexes i, j are state labels, $P_i^{eq} = n_i/M$ is the equilibrium probability of the state i (n_i snapshots over a total number of M) and $P_{ij}^{eq}(\tau) = n_{ij}(\tau)/(M-1)$ is the probability flux for the transition $i \to j$ at the lag time τ, where $n_{ij}(\tau)$ is the total number of transitions $i \to j$. All the quantities are estimated within the lag time τ of 20 ps, which is the saving time of the trajectories. To test the Markov property of the time series at the lag time τ a non-Markovian flux was estimated by comparing the one-step transition matrix $T_{jk}(\tau)$ with the two-step transition matrix $T_{ijk}(\tau)$ for the transition $i \to j \to k$. The two-step transition matrix is

$$T_{ijk}(\tau) = P_{ijk}^{eq}(\tau)/P_{ij}^{eq} \simeq n_{ijk}(\tau)/n_{ij}(\tau) \qquad (2)$$

where $P_{ijk}^{eq}(\tau)$ and $n_{ijk}(\tau)$ are respectively the probability flux and the total number of transitions $i \to j \to k$. The Markov property is valid if the identity $T_{ijk}(\tau) = T_{jk}(\tau)$ is satisfied for any i. Using the relation (2) and summing up over all the two-step transitions one obtains the total non-Markovian flux

$$F(\tau) = 1 - \sum_{i \to j \to k} P_i^{eq} T_{ij}(\tau) T_{jk}(\tau) \qquad (3)$$

The non-Markovian flux is a probability flux which reflects the overall error made by assuming the Markov approximation on a time series at a certain lag time τ. The statistical significance of the clusters plays an important role if one is interested to describe a time series adopting a Markov approximation. A procedure based on the reassignment of the clusters memberships is employed here to achieve the Markovianity of the time series: the snapshots of the low-populated clusters are reassigned to the statistically significant clusters according to their causal connectivity along the time series. This is attained by reprocessing the time series of clusters to obtain a time series of "causally grouped mesostates": when a snapshot of an insignificant cluster (size $<$ cutoff) is encountered, it is causally reassigned to the next significant cluster (size $\geq$ cutoff). The cutoff is chosen such that the resulting time series are Markovian, or more precisely, have a non-Markovian flux less than 1%. For the present simulation of protein ssG, 200 causally grouped mesostates resulted from a cluster size cutoff of 250 snapshots (see Figure 2 in Suppl. Mat). The simplicity of the procedure is rooted on the hypothesis that the dynamics of the polypeptide takes place only between stable states where the system can partially diffuse loosing memory of previously explored states. Remarkably, at the lag time of 20 ps the overall error of the Markov approximation is less than 1% for the 200 causally grouped mesostates, while it is 7.5% if one considers for the transition matrix the 3124 clusters with two or more snapshots (see Figure 2 in Suppl. Mat). The difference justifies the adoption of the causally grouped mesostates for the Markov approximation. Thus, once a time series of causally grouped mesostates is provided, the transition matrix $T_{ij}(\tau)$ can be estimated, where now the indexes i, j run from 1 to 200.

In order to show that the validity of the Markov approximation at lag time $\tau = 20$ ps is good enough for longer time scales extrapolations, transition matrices for longer lag times (from 20 ps to 20 ns) were estimated from

the causal grouped time series. The relaxation times corresponding to the eigenvalues show a robustness in the values of the slower relaxation times (see Figure 3 in Supp. Mat.) within these time ranges. Moreover, the distribution of the first passage times to the folded states calculated from MD and using the Markov approximation compare very well in both shape and time scales, indicating a substantial equivalence in the kinetics of the original and the modeled processes (see Figure 4 in Supp. Mat.). Both these two results suggest that the Markov approximation adopted for the causal grouped mesostates at 20 ps of lag time is robust enough to infer the long time kinetics of the folding process.

The equilibrium counterpart of the transition matrix $\mathbf{T}(\tau)$ is the matrix of mean first passage times (MFPT) $\mathbf{M}$ whose entries M_{ij} give the mean hitting time for the transitions between the mesostates $i \to j$, averaged over all the possible connecting pathways. By assuming the ergodicity of the underlying finite Markov chain the M_{ij} matrix is given by a system of linear equations such as

$$\begin{aligned} M_{ij} &= \tau + \sum_{k \neq j} T_{ik}(\tau) M_{kj} \\ M_{ii} &= \sum_{k} T_{ik}(\tau)(M_{ki} + \tau) \end{aligned} \tag{4}$$

that are exactly solvable when the number of states is small (28). Assigning the index 1 to the folded mesostate, then the first column of the MFPT matrix (M_{i1}) gives the mean folding times from individual mesostates to the folded one. To facilitate the reading of the $\mathbf{M}$ matrix, the indexes were reordered in such a way that the low numbers (from 1) are the mesostates with small folding times, while large numbers (up to 200) have longer folding times. Thus, the first row of the $\mathbf{M}$ matrix satisfies the inequalities $M_{1\,1} \leqslant M_{2\,1} \leqslant \cdots \leqslant M_{200\,1}$. The indexes of the reordered MFPT matrix are adopted for the labeling of the mesostates throughout this work.

E. Static and dynamic correlations of secondary structure

The time series of SSS[8] allows the adoption of information theory methods to investigate the underlying structural mechanisms of folding. For each residue a probability $\pi_i(s)$ can be defined where i is the residue number and s is one of the eight secondary structure symbols. Similarly, a pairwise probability $\pi_{ij}(ss')$ is defined between two residues i and j, and secondary structure s and s'. Both probabilities are estimated from the time series of SSS[8]. A static correlation between pairs of residues can be evaluated from the ensemble of visited strings by calculating a pairwise mutual information. In information theory the mutual information between two random variable measures their mutual dependence (29). With the probabilities previously defined the mutual information between two residues is defined as

$$I_{ij} = \frac{1}{\ln 8} \sum_{ss'} \pi_{ij}(ss') \ln \frac{\pi_{ij}(ss')}{\pi_i(s)\pi_j(s')} \tag{5}$$

which is a normalized quantity that is zero when the residues i and j are totally uncorrelated, and 1 when they are totally correlated.

The static mutual information can be generalized to obtain a correlation function with the aim to evaluate the dynamics of formation of secondary structure. We define a time dependent pairwise probability $\pi_{ij}(ss', t)$ that two residues i, j assume secondary structure ss' at the time t. A time dependent mutual information is defined as

$$I_{ij}(t) = \frac{1}{\ln 8} \sum_{ss'} \pi_{ij}(ss', t) \ln \frac{\pi_{ij}(ss', t)}{\pi_i(s)\pi_j(s')} \tag{6}$$

from which the pairwise normalized correlation function between two residues reads

$$C_{ij}(t) = \frac{I_{ij}(t) - I_{ij}(\infty)}{I_{ij}(0) - I_{ij}(\infty)} \tag{7}$$

where $I_{ij}(\infty)$ and $I_{ij}(0)$ are the equilibrium and the static values of the mutual information, respectively.

III. RESULTS AND DISCUSSION

All analyses are based on a 15-μs molecular dynamics simulation of protein ssG at 330 K started from a fully extended conformation with the backbone dihedral angles equal to 180 degrees. First the 750000 snapshots (saved every

20 ps) were clustered by C_α RMSD. From the resulting 132006 clusters the causal grouping procedure generated 200 mesostates (see Methods section). The most populated mesostate contains 3.5% of the snapshots (Table II) and corresponds to the native topology of protein G.

A. Fast folding to a molten globule

Multiple folding and unfolding events are sampled along the 15-μs trajectory as illustrated by the time series of C_α root mean square deviation (RMSD) from the X-ray structure (PDB code 1pgb) and the fraction of native contacts (Figure 1). Note that the term folding is used here in a relaxed sense to indicate that the molten-globule state with native topology has been reached. In fact, in simulation segments where the conformation has the native topology, the C_α RMSD oscillates between 2.5 Å and 5 Å from the X-ray structure, the radius of gyration varies between 9 Å and 11 Å, and the fraction of native contacts between 0.6 and 0.9. These range of values reflect a fluid-like behavior typical of a molten globule. Such behavior emerges also from the structural overlap of the conformations in the most populated mesostate (Figure 2A). More quantitatively, the average value of the pairwise C_α RMSD within this mesostate is 3.5 Å. Interestingly, within the most populated mesostate the largest structural variability is observed at loops L1, L3, and L4 (Figure 2A), in agreement with the largest deviations between X-ray structure (30) and NMR conformers (31; 32).

As a basis of comparison, using the same temperature, three 1-μs simulations of the wild-type sequence started from extended got trapped into compact non-native conformations with a C_α RMSD from the X-ray structure ranging from 7 to 14 Å. Note also that in control simulations started from the folded state the wild-type protein is structurally stable on a 1-μs time scale.

B. Heterogeneous denatured state

The network representation of the 200 causal mesostates (nodes) and their transition matrix (links) illustrates the configuration space of protein ssG (Figure 3). A semiquantitative description of the free energy basins emerges from the thickness of the links and size of the nodes, which reflect the probabilities of internode transition and node population, respectively. Moreover, the quality-threshold algorithm is used to partition the network into basins, which are emphasized by different color in Figure 3. Note that the network of causal mesostates is more informative than the original conformational space network (22), which depicted only the dynamic connectivity but did not show quantitative information on kinetics. The basin of the folded mesostate includes also other mesostates with the secondary structure of protein G, and has a population of 21.7% (red basin in Figure 3). Although its most populated mesostate has the correct protein G topology, it contains other mesostates with one hairpin flipped (mesostate 35 in Figure 3). These mesostates with slightly different topology interconvert very rapidly within the most populated basin. The mesostates in the folded basin are stabilized mainly by enthalpy (Table II). In particular, the most populated mesostate has an average effective energy 12.4 kcal/mol more favorable than the effective energy averaged over the entire trajectory. The most populated basin is in fast exchange with a basin (of statistical weight of 6.3%) that contains mesostates having both hairpins flipped with respect to the native topology of protein G (mesostate 49 and green basin in Figure 3).

The unfolded state is heterogeneous and is made up of mesostates with different relative amount of α-helical and β-sheet content. The three-helix bundle mesostates 133 and 147 (gray in Figure 3) connect two unfolded basins with a mixture of α-helical and β-sheet content. One of these two basins has statistical weight of 10.3% (cyan in Figure 3) and includes conformations with a three-stranded β-sheet packed against a long helix (mesostate 164), while the other has a weight of 13.1% (violet in Figure 3) and includes mesostates with two long helices and a short β-hairpin (mesostate 119). Notably, at the border of the network there are several mesostates with a very high β-sheet content (e.g., mesostates 66, 91, 198, and 200). They can be considered off-pathway traps because the main folding transitions connect the unfolded basins consisting of conformations with mixed secondary struc-

ture content to the folded basin (see next subsection).

C. Folding mechanisms I

The distribution of the first passage times to reach the folded mesostate, calculated on the time series of 200 causally grouped mesostate, is a single exponential curve with a mean folding time of 163 ns (see Figure 4 in Suppl. Mat.). This apparent simplicity is in striking contrast with the complexity of the transition-matrix network (Figure 3). As explained in the Methods section the equilibrium extrapolation of the Markov chain is the matrix of MFPT values, which gives the equilibrium transition time between pairs of states. The graphical rendering of the MFPT matrix shows in a compact way the kinetic distance between all pairs of causal mesostates (Figure 4). The band structure of the MFPT matrix provides useful informations on the folding mechanism of the ssG protein. The horizontal bands are due to the fact that the MFPT matrix is a directed matrix, so that the mean time to go from a mesostate i to j is different than for the inverse transition, because different are in general the corresponding pathways. The bands give the overall kinetic accessibility of individual mesostates. There are four rather distinct kinetic regions of the conformation space. Mesostates 1-60 rapidly exchange with the folded mesostate and can be accessed from all other mesostates within 100-300 ns. Mesostates 61-104 are transient and most of them separate the folded region from the unfolded basins. In the region 105-175 are located most of the unfolded basins (α/β and only α structures), while the fourth region, mesostates 176-200, includes the kinetic traps with high β-sheet content.

D. Folding mechanisms II

The secondary structure formation is analyzed by means of pairwise correlations whose calculation is based on the mutual information between pairs of residues (see Methods). Both static and dynamic correlations are calculated for all residue pairs. The static correlation is evaluated by calculating the normalized mutual information between pairs of residues on the ensemble of non redundant strings of secondary structure observed in the simulation of ssG protein (Figure 5). The modular pattern of the matrix suggests that the interactions responsible for the secondary structure formation are present mainly between the homopolymer segments of the protein. The highest correlations is observed for the local secondary structure, in particular the residues involved in the α-helix and the two native β-hairpins (correlation $\gtrsim 20\%$). Long range correlations define all possible tertiary topologies corresponding to a four-stranded β-sheet packed on a helix. These correlations are weaker than the local ones. Their averaged values are $\sim$4% for S1S4, $\sim$ 3% for both S1S3/S2S4 and $\sim$ 1% for S2S3. Notice that the S1S4 correlation corresponds to the β-strand arrangement as in the correct protein G topology. The long range correlations S2-H and H-S3 are weaker than those mentioned above, and give rise to a long helix involving residues Thr_{12}-Ala_{37} or Ala_{23}-Thr_{47}, respectively. Overall, the static correlations indicate that there is a propensity of protein ssG to assume the very same secondary structure of protein G.

Dynamic correlations provide a mechanistic view on what are the sequential events taking place in secondary structure formation. The correlations are evaluated by calculating the mutual information between pairs of residues as a function of time and then averaging within the defined fragments (see Methods). The times at which the dynamic correlation reaches a value of 0.5 for the α-helix and the C-terminal β-hairpin S3S4 are similar (about 5 ns), while those for the N-terminal β-hairpin S1S2 and the parallel arrangement of S1S4 are about 10 ns and 15 ns, respectively (Figure 6). All other combinations of β-strands, which yield non-native topologies, have slower correlations time, suggesting a sequence of events for folding which is compatible with a diffusion-collision mechanism (33; 34). According to such mechanism individual elements of secondary structure (the α-helix, S1S2, or S3S4) can form independently from each other. Interactions among segments that are distant along the sequence, (e.g., native S1S4, and non-native S1S3 or S2S4) promote the formation of a complex tertiary structure by coalescence.

IV. CONCLUSIONS

To investigate a putatively primordial protein we have dramatically simplified the sequence of protein G using only three types of residues: Glycine, alanine, and threonine. Molecular dynamics simulations of the simplified-sequence variant of protein G (termed ssG) provide strong evidence that a heteropolymer with a limited assortment of monomer types is able to adopt a complex topology. In fact, reversible folding to the wild-type native topology has been achieved in this work by using a force field-based (i.e., transferable) potential. On the other hand, using the same force field and simulation protocol the wild-type sequence of protein G does not fold on the same time scale. [Note that structured peptides (α-helices and β-sheets) fold to the correct conformation with the very same force field and implicit solvent model as documented in previous simulation studies (18; 19; 25; 35; 36).]

The Markov-chain analysis of the atomistic simulations of protein ssG was used to investigate the unfolded state and folding mechanism, which is not possible by conventional experimental techniques. Three main results emerge from this analysis. First, rapid folding is observed for a simplified-sequence variant of a protein with α/β topology. Note that this topology is more heterogeneous than the all-β topology of wild type and simplified variant of protein SH3 (13). The MFPT prediction from Markov approximation also indicates that the lack of diversity of interactions reduces the frustration of the free-energy landscape so that conformations with significantly different content of secondary structure interconvert very rapidly. The correlation analysis for secondary structure formation suggests that the molten-globule state is reached through multiple pathways (37) and by a diffusion-collision mechanism (framework) (33; 34) which is due to the strong secondary structure propensity of the helical segment and the two β-hairpins. In fact, the initial folding events are the independent formation of the local elements of secondary structure. The assembly of regular elements of secondary structure takes place by coalescence and is mainly driven by backbone-backbone hydrogen bonding. The reduced side chain heterogeneity allows the system to explore a large variety of topologies that are compatible with the secondary structure of protein G.

Second, reduced alphabets of amino acids seem to be suitable to define globular folds with abundant secondary structure elements but they do not encode for the specificity of tertiary contacts required for a native, i.e., functional, structure. However, low complexity alphabets of amino acids have been shown recently to be suitable for molten globular active enzymes (38; 39). Furthermore, simplified sequences of a three-helix bundle fold (protein G_A88) and an α/β fold (protein G_B88, which is the very same domain of protein G used in our simulations) with 88% sequence identity were shown to possess different structure and function (40). Therefore, the information determining the fold seems to be "higly concentrated in a few amino acids" (40), i.e., only 7 of 56, and very recent results by the same authors indicate only 3 of 56 (41). Our simulation results, in particular the variety of topologies observed for protein ssG (which include the folds of both protein G_A88 and G_B88), provide the following explanation of the experimental findings: It is likely that both folds are populated by both G_A88 and G_B88, but only one fold, the statistically predominant one, is observed in the ensemble experiments. Moreover, the relative statistical weight can be easily shifted towards a particular fold by changing only a small subset of the residues.

Third, despite the reduced diversity in the interactions the denatured state is heterogeneous as it consists of structures with a secondary structure content ranging from fully α-helical to fully β-sheet. The latter are kinetic traps and might promote aggregation. Interestingly, Langevin dynamics simulations with a coarse-grained model of an amphipathic polypeptide indicate that a minor increase (≤ 1 kcal/mol) in relative stability of a β-aggregation prone state, can result in a dramatic acceleration of fibril formation rates (42; 43). On the experimental side, protein G (more precisely the same domain of protein G as in the present study) was shown to form amyloid fibrils under mild denaturation conditions (44). Furthermore, several double-mutants with reduced thermodynamic stability were observed to aggregate with high reproducibility in the same study. In other words, by

controlling the stability of the protein, through mutations or variation of the experimental conditions, it was possible to modulate the ability to form fibrils. Notably, the key requirement for fibril formation was to choose conditions in which the population of intermediate states present during the unfolding transition was maximized. Furthermore, by comparing mutations at different strands of protein G the same authors have provided evidence that the overall stability of protein G is the key determinant for amyloid formation and not the specific location of destabilizing mutations (45).

On the basis of the experimental data on protein G amyloid-fibril formation and the present simulation results, we suggest that the enrichment of a primordial (i.e., reduced) alphabet of residues has been directed by evolution towards a double purpose: the optimization of protein function (which in most cases requires a stable folded structure) and at the same time the elimination of non-native conformations that are aggregation-prone by means of frustration and competing interactions. Dramatically reduced alphabets of amino acids are suitable to define elementary folds but they do not encode the sufficient complexity such that both these optimization prescriptions can be achieved by evolution.

We would like to conclude by quoting from a paper by F. Crick of exactly 40 years ago (46) (which we discovered while finalizing this manuscript): *"It certainly seems unlikely that all the present amino acids were easily available at the time the code started. Certainly tryptophan and methionine look like later additions. Exactly which amino acids were then common is not yet clear, though most lists would include glycine, alanine, serine and aspartic acid."* Without knowing it, the simplified three-letter alphabet used in the present simulation study included two of these four residues and threonine (which is similar to serine). Furthermore, glycine and alanine were first observed (together with aspartic acid) in the remarkable experiment of S. Miller (47) on the amino acid synthesis under primitive conditions.

Acknowledgments

We thank Andrea Cavalli for interesting discussions. The simulations were performed on the Matterhorn cluster of the University of Zurich and we thank C. Bolliger and Dr. Godknecht for computer support.

References

[1] Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. Science 254:1598–603.

[2] Chan H, Dill K (1998) Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. Proteins: Structure, Function, and Bioinformatics 30:2–33.

[3] Karplus M (2000) Aspects of protein reaction dynamics: deviations from simple behavior. J Phys Chem B 104:11–27.

[4] Wallin S, Shakhnovich E (2008) Understanding ensemble protein folding at atomic detail. Journal of Physics: Condensed Matter 20:283101.

[5] Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230.

[6] Shakhnovich EI (1998) Protein design: a perspective from simple tractable models. Folding & Design 3:R45–R58.

[7] Shakhnovich E (2006) Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 106:1559–1588. doi:10.1021/cr040425u.

[8] Sali A, Shakhnovich E, Karplus M (1994) How does a protein fold? Nature 369:248–251.

[9] Kaya H, Chan HS (2000) Energetic components of cooperative protein folding. Phys Rev Lett 85:4823–4826.

[10] Davidson AR, Sauer RT (1994) Folded proteins occur frequently in libraries of random amino acid sequences. Proc Natl Acad Sci USA 91:2146–2150.

[11] Davidson AR, Lumb KJ, Sauer RT (1995) Cooperatively folded proteins in random sequence libraries. Nature Struct Biol 2:856–864.

[12] Cordes MH, Davidson AR, Sauer RT (1996) Sequence space, folding and protein design. Curr Opin Struct Biol 6:3–10.

[13] Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, et al. (1997) Functional rapidly folding proteins from simplified amino acid sequences. Nature Struct Biol 4:805–809.

[14] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, et al. (1983) CHARMM - a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217.

[15] Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A (2007) Wordom: a program for efficient analysis of molecular dynamics simulations. Bioinformatics 23:2625–2627.

[16] Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. J Chem Phys 105:1902–1921.

[17] Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins: Structure, Function, and Bioinformatics 46:24–33.

[18] Ferrara P, Apostolakis J, Caflisch A (2000) Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. J Phys Chem B 104:5000–5010.

[19] Settanni G, Rao F, Caflisch A (2005) Φ-Value analysis by molecular dynamics simulations of reversible folding. Proc Natl Acad Sci USA 102:628–633.

[20] Eaton WA, Munoz V, Hagen SJ, Jas GS, Lapidus LJ, et al. (2000) Fast kinetics and mechanisms in protein folding. Ann Rev Biophys Biomol Struct 29:327–359. doi:10.1146/annurev.biophys.29.1.327.

[21] De Alba E, Santoro J, Rico M, Jiménez MA (1999) De novo design of a monomeric three-stranded antiparallel beta-sheet. Prot Sci 8:854–865.

[22] Rao F, Caflisch A (2004) The protein folding network. J Mol Biol 342:299–306.

[23] Krivov SV, Karplus M (2004) Hidden complexity of free energy surfaces for peptide (protein) folding. Proc Natl Acad Sci USA 101:14766–14770.

[24] Hubner IA, Deeds EJ, Shakhnovich EI (2006) Understanding ensemble protein folding at atomic detail. Proc Natl Acad Sci USA 103:17747–17752.

[25] Ihalainen J, Paoli B, Muff S, Backus E, Bredenbeck J, et al. (2008) Alpha-Helix folding in the presence of structural constraints. Proc Natl Acad Sci US A 105:9588–93.

[26] Heyer L, Kruglyak S, Yooseph S (1999) Exploring Expression Data: Identification and Analysis of Coexpressed Genes. Genome Research 9:1106.

[27] Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.

[28] Kemeny J, Snell J (1976) Finite Markov Chains. Springer.

[29] Cover T, Thomas J (1991) Elements of information theory. Wiley-Interscience New York, NY, USA, 542 pp.

[30] Gallagher T, Alexander P, Bryan P, Gilliland GL (1994) Two crystal structures of the b1 immunoglobulin-binding domain of streptococcal protein g and comparison with nmr. Biochemistry 33:4721–4729.

[31] Gronenborn AM, Frank MK, Clore GM (1996) Core mutants of the immunoglobulin binding domain of streptococcal protein g: stability and structural integrity. FEBS Lett 398:312–316.

[32] Lian LY, Derrick JP, Sutcliffe MJ, Yang JC, Roberts GC (1992) Determination of the solution structures of domains ii and iii of protein g from streptococcus by 1h nuclear magnetic resonance. J Mol Biol 228:1219–1234.

[33] Karplus M, Weaver DL (1976) Protein-folding dynamics. Nature 260:404–6.

[34] Islam SA, Karplus M, Weaver DL (2004) The role of sequence and structure in protein folding kinetics; the diffusion-collision model applied to proteins l and g. Structure 12:1833–45. doi:10.1016/j.str.2004.06.024.

[35] Hiltpold A, Ferrara P, Gsponer J, Caflish A (2000) Free energy calculation of the helical peptide Y(MEARA)6. J Phys Chem B 104:10080–10086.

[36] Muff S, Caflisch A (2008) Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. Proteins 70:1185–1195. doi:10.1002/prot.21565.

[37] Wright C, Lindorff-Larsen K, Randles L, Clarke J (2003) Parallel protein-unfolding

pathways revealed and mapped. Nat Struct Biol 10:658–62.

[38] Walter KU, Vamvaca K, Hilvert D (2005) An active enzyme constructed from a 9-amino acid alphabet. J Biol Chem 280:37742–37746. doi:10.1074/jbc.M507210200.

[39] Vamvaca K, Vögeli B, Kast P, Pervushin K, Hilvert D (2004) An enzymatic molten globule: efficient coupling of folding and catalysis. Proc Natl Acad Sci USA 101:12860–12864. doi:10.1073/pnas.0404109101.

[40] Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2007) The design and characterization of two proteins with 88structure and function. P Natl Acad Sci Usa 104:11963–8. doi:10.1073/pnas.0700922104.

[41] He Y, Chen Y, Alexander P, Bryan PN, Orban J (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. Proc Natl Acad Sci USA 105:14412–14417. doi:10.1073/pnas.0805857105.

[42] Pellarin R, Caflisch A (2006) Interpreting the aggregation kinetics of amyloid peptides. J Mol Biol 360:882–892. doi:10.1016/j.jmb.2006.05.033.

[43] Pellarin R, Guarnera E, Caflisch A (2007) Pathways and intermediates of amyloid fibril formation. J Mol Biol 379:917–924. doi:10.1016/j.jmb.2007.09.090.

[44] Ramirez-Alvarado M, Merkel J, Regan L (2000) A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. Proc Natl Acad Sci USA 97:8979–8984.

[45] Ramirez-Alvarado M, Regan L (2002) Does the location of a mutation determine the ability to form amyloid fibrils? J Mol Biol 323:17–22. doi:10.1016/S0022-2836(02)008490-9.

[46] Crick FHC (1968) The origin of the genetic code. J Mol Biol 38:367–379.

[47] Miller S (1953) A Production of Amino Acids Under Possible Primitive Earth Conditions. Science 117:528–529.

[48] Andersen CAF, Palmer AG, Brunak S, Rost B (2002) Continuum secondary structure captures protein flexibility. Structure 10:175–184.

[49] Auber D (2003) Tulip : A huge graph visualisation framework. In: Mutzel P, Jünger M, editors, Graph Drawing Softwares, Springer-Verlag, Mathematics and Visualization. pp. 105–126.

12

Tables

Sequences:	
protein G	MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
protein ssG	TTTTTTTTTGGTTTTTTTTTGGAAAAAAAAAAAAAAAAGGTTTTTTTTGGTTTTTTT
Secondary structure string:	
	-EEEEEEEESSEEEEEEE-SSHHHHHHHHHHHHHHH----EEEEETTT-EEEEE-

TABLE I Sequences of proteins G and ssG. The secondary structure string was determined using the X-ray structure (30). In the DSSP string the letters E, H, S, T, and "-", correspond to extended, α-helical, bend, hydrogen-bonded turn, and unstructured, respectively (48).

13

Rank[a]	P_i [%]	ΔG_i [kcal/mol]	ΔE_i [kcal/mol]	$-T\Delta S_i$ [kcal/mol]	M_{i1} [ns]	α-helix [%]	β-sheet [%]
1	3.5	-1.0	-12.4	11.4	1	25	44
49	2.7	-0.9	-4.8	3.9	11	24	41
127	2.5	-0.8	-2.5	1.7	90	64	4
147	2.1	-0.7	3.0	-3.7	95	57	5
133	1.8	-0.6	3.7	-4.3	92	51	9
128	1.8	-0.6	3.9	-4.5	90	53	8
35	1.6	-0.5	-8.8	8.3	9	26	44
186	1.6	-0.5	0.9	-1.4	101	64	3
183	1.6	-0.5	2.9	-3.4	98	53	10
16	1.6	-0.5	-4.8	4.3	4	29	38
119	1.6	-0.5	-7.9	7.4	87	55	13
182	1.5	-0.5	1.6	-2.1	98	67	3
134	1.4	-0.4	4.3	-4.7	92	52	8
153	1.3	-0.4	-0.2	-0.2	96	63	4
125	1.3	-0.4	1.6	-2.0	89	53	10
164	1.2	-0.3	-6.2	5.9	96	42	29
139	1.1	-0.3	6.6	-6.9	94	38	16
123	1.1	-0.3	1.9	-2.2	89	53	10
24	1.0	-0.2	2.2	-2.4	6	35	27
179	1.0	-0.2	0.5	-0.7	97	43	21
174	1.0	-0.2	6.1	-6.3	97	40	15
171	1.0	-0.2	7.3	-7.5	96	39	19
152	1.0	-0.2	6.7	-6.9	96	43	13
138	1.0	-0.2	6.0	-6.2	94	32	24
105	1.0	-0.2	3.3	-3.5	83	47	14
48	0.9	-0.1	-5.9	5.8	11	22	44
4	0.9	-0.1	-10.4	10.3	2	25	37
200	0.9	-0.2	-1.4	1.2	314	0	74
198	0.9	-0.1	-2.7	2.6	201	2	60
172	0.9	-0.1	7.7	-7.8	97	31	22
132	0.9	-0.1	1.2	-1.3	92	31	31
129	0.9	-0.1	2.0	-2.1	90	46	15
121	0.9	-0.1	0.0	-0.1	88	32	32
116	0.9	-0.1	2.9	-3.0	87	51	9
10	0.9	-0.1	-5.9	5.8	3	28	39
91	0.8	-0.1	4.8	-4.9	73	12	45
87	0.8	-0.0	1.8	-1.8	68	31	30
75	0.8	-0.1	8.4	-8.5	38	34	20
21	0.8	-0.1	1.9	-2.0	5	28	31
184	0.8	-0.0	0.6	-0.6	99	41	23
161	0.8	-0.1	-1.0	0.9	96	27	37
76	0.7	-0.0	4.7	-4.7	43	32	21
47	0.7	0.0	-0.5	0.5	11	25	38
29	0.7	0.0	-0.6	0.6	7	32	28
162	0.7	0.1	-6.5	6.6	96	43	26
151	0.7	0.0	3.7	-3.7	96	39	20
137	0.7	0.0	0.6	-0.6	93	26	34
124	0.7	0.1	-1.6	1.7	89	59	7
118	0.7	0.1	0.4	-0.3	87	47	16
113	0.7	0.0	7.9	-7.9	87	39	13

TABLE II The 50 most populated causally grouped mesostates. [a]The rank originates from sorting the 200 mesostates according to the folding times M_{i1} calculated by the equilibrium evolutions of the Markov chain. Structures in mesostates with rank in boldface are shown in Figures 3 and 4. Average effective energy (sum of force field and SAS solvation energy) relative to the whole simulation $\Delta E_i = \langle E_i \rangle - \langle E \rangle$, where the $\langle E_i \rangle$ and $\langle E \rangle$ values are calculated over the snapshots in the causally grouped mesostate i and the whole trajectory, respectively. Note that, in any force field, the absolute value of the effective energy is arbitrary and only ΔE values relative to a reference state are meaningful. The free energy differences are calculated by the relation $\Delta G_i = -k_B T \sum_j P_j \ln(P_i/P_j)$. Consequently, the entropy contribution to the free energy difference $-T\Delta S_i$ is calculated using the relation $-T\Delta S_i = \Delta G_i - \Delta E_i$.

Figures

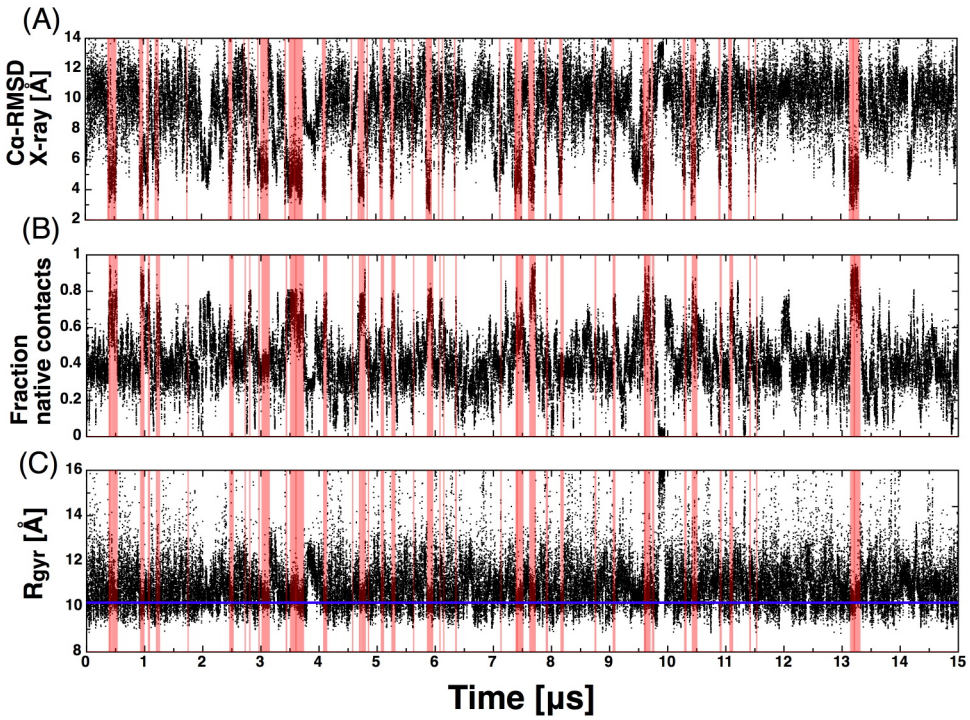

FIG. 1 Rapid and reversible folding of protein ssG. Folding events along the time series are emphasized by pink vertical stripes. (A) Time series of the C_α RMSD from the X-ray structure (PDB code 1pgb). The two N-terminal and two C-terminal residues were excluded from the RMSD calculation. (B) Time series of the fraction of native contacts in the backbone. The native contacts were defined using the X-ray structure and considering the heavy atoms in the backbone for residues that are ≥ 3 distant along the sequence. A contact exists when the distance is smaller than 7 Å, which yields 422 native contacts in the X-ray structure. (C) Time series of the radius of gyration with the blue line corresponding to the native radius of gyration of protein G ($R_{\rm gyr} = 10.2$ Å). The mean first passage time to reach the folded mesostates, calculated on the time series, is 163 ± 157 ns (see Figure 4 in Suppl. Mat).

15

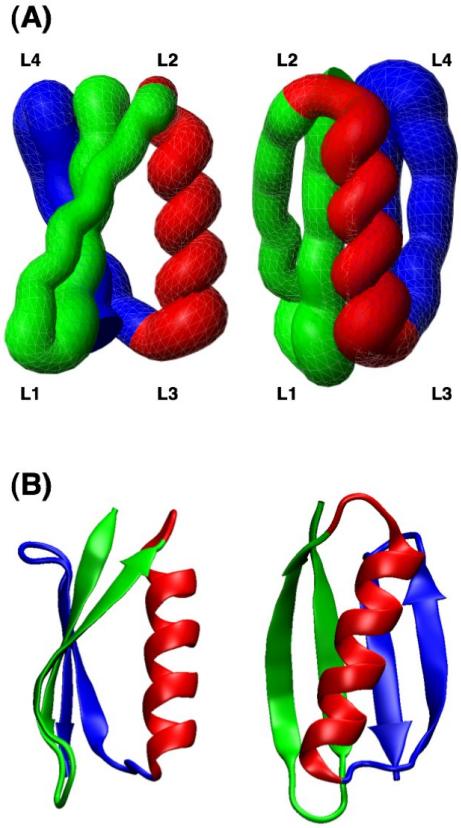

FIG. 2 Comparison of the molten-globule state extracted from the simulations of protein ssG (A) and the X-ray structure of protein G (B). The N-terminal β-hairpin, central α-helix, and C-terminal β-hairpin are in green, red, and blue, respectively. The tube-like rendering in (A) was generated using 100 snapshots from the most populated mesostate. Note that the topology of protein ssG is the same as the one of the wild-type protein but the lack of long side chains and specific contacts in the former results in a flatter β-sheet and a slightly different orientation of the α-helix with respect to the β-sheet.

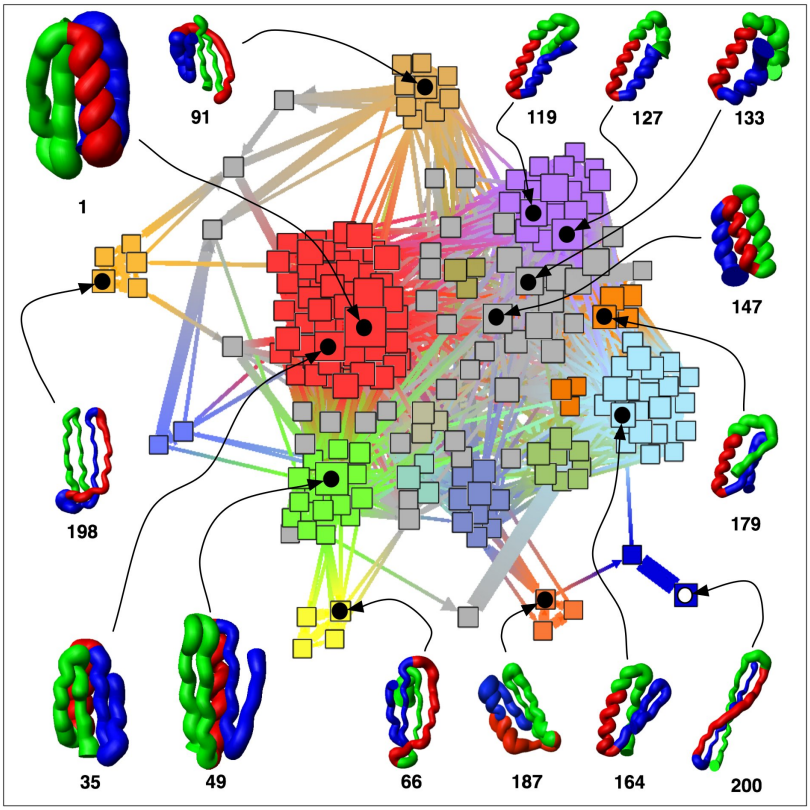

FIG. 3 The network representation of the transition matrix. The tube-like rendering of representative conformations was generated as in Figure 2A. The nodes are the 200 mesostates determined by causal grouping while the links are the transition probabilities T_{ij} extracted from the trajectory. The size of the nodes is proportional to their population, while the size of the links reflects the probability value in the transition matrix with a lag time of 20 ps. The position of the nodes in the network was determined by the spring-embedder visualization algorithm of the program Tulip (49), which takes into account the values of the transition matrix to optimize the node positioning in the plane. The color of the nodes is assigned according to basin's membership, which is determined by clustering the transition matrix of the 200 mesostates using the quality-threshold algorithm with a cutoff of $T_{ij} > 0.0001$. Color assignment begins from the node that has the largest number of neighbors with link value, i.e., transition probability, above the cutoff. With this procedure, 52 basins were identified and the most populated includes the folded mesostate. Of these 52 basins, 28 and 9 consist of only 1 and 2 mesostates, respectively (gray nodes). Yet, the total weight in 1-mesostate and 2-mesostate basins is only 18% and 9%, respectively.

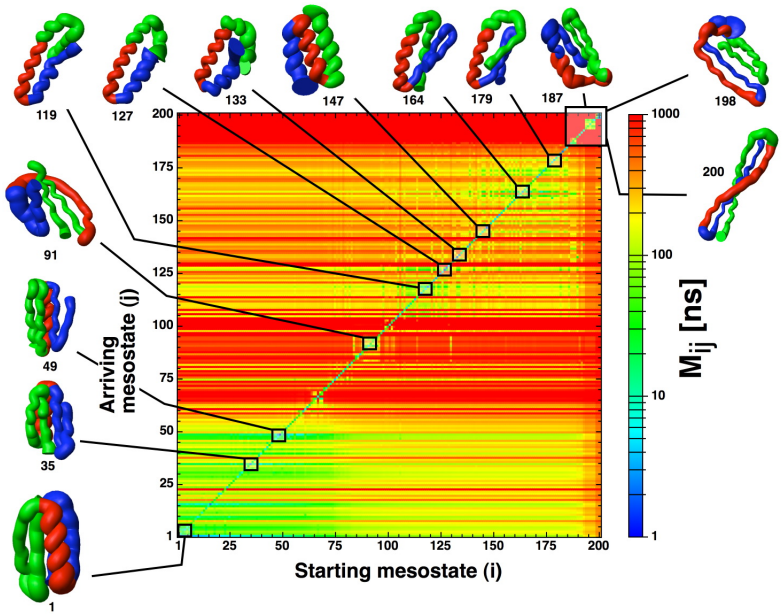

FIG. 4 Folding kinetics illustrated by the reordered MFPT matrix M_{ij} of the 200 causally grouped mesostates. An element of the matrix is the MFPT for the $i \rightarrow j$ transition at equilibrium. Note that the matrix is not symmetric because each entry is an MFPT value and not a flux. The latter is the MFPT value multiplied by the equilibrium probability and would yield a symmetric matrix. Horizontal rows are equilibrium transitions from all the mesostates i (x axis) to a specific j (y axis). The indices (i,j) are ordered from 1 (fastest relaxation to the most populated mesostate, which belongs to the molten-globule state with native topology) to 200 (slowest relaxation). The green-yellow band in the bottom indicates that the native-like molten-globule state can be reached rapidly from all other mesostates. The conformations with high β-sheet content are kinetically most distant from the most populated mesostate. The mesostates with helical bundles and/or mixed α and β content interconvert rapidly.

18

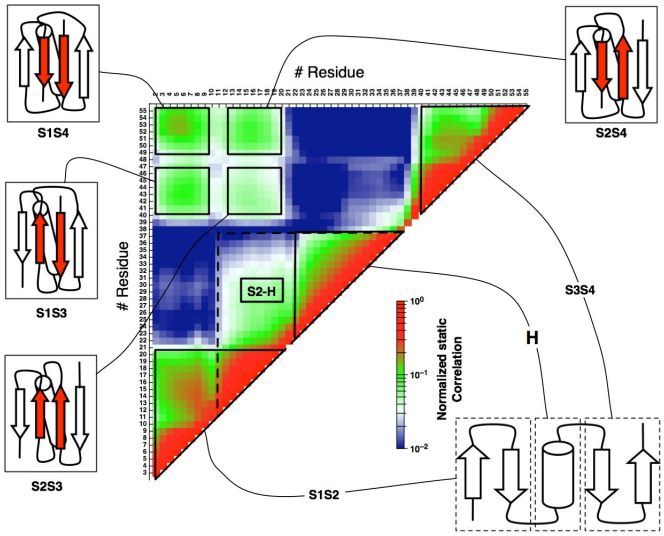

FIG. 5 Matrix of the static correlation of secondary structure I_{ij} (Eq. 5). The modular pattern suggests that the interactions responsible for secondary structure formation are present between the homopolymer segments of the protein ssG. The cartoons are shown to illustrate the secondary structure elements having the highest correlations. Abbreviations: H=Ala_{23}-Ala_{37} for the poly-Ala and S1=Thr_1-Thr_9, S2=Thr_{12}-Thr_{20}, S3=Thr_{40}-Thr_{47}, and S4=Thr_{50}-Thr_{56} for the poly-Thr.

19

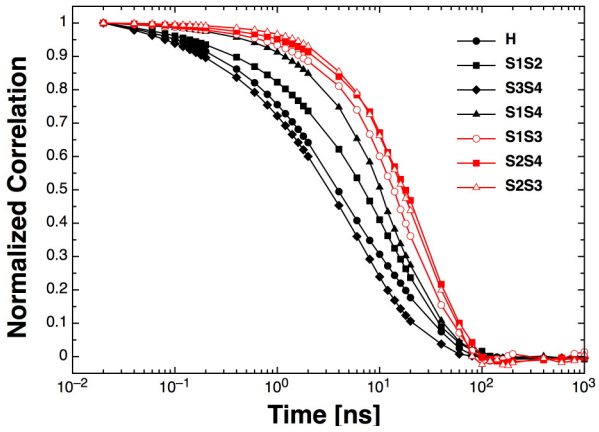

FIG. 6 Dynamic correlation between secondary structure elements C_{ij} (Eq. 7). Native and non-native elements of secondary structure are in black and red, respectively. Different time scales for secondary structure formation suggest a folding mechanism compatible with the framework model. The curve H represents the autocorrelation within the poly-Ala α-helix, while S1S2 (N-terminal β-hairpin), S3S4 (C-terminal β-hairpin), S1S4 (N/C-terminal two-stranded parallel β-sheet), as well as the non-native arrangements S1S3, S2S4, and S2S3 reflect the association of poly-Thr β-strands.

How does a simplified-sequence protein fold? SUPPLEMENTARY MATERIAL

Enrico Guarnera, Riccardo Pellarin, and Amedeo Caflisch*

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
FAX 0041 44 635 68 62

(Dated: February 9, 2009)

I. CLUSTERING

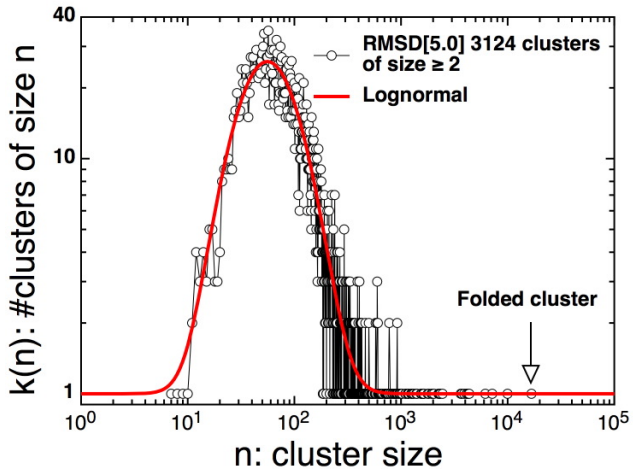

FIG. 1 Statistical significance of the clusters. The 5Å cutoff in C_α-RMSD and the quality-threshold algorithm used for clustering yielded 23% of unassigned conformers. A total of 3124 clusters were found with size ≥ 2. The distribution follows a lognormal profile. On the right tail for $n \gtrsim 100$ are the statistically significant clusters.

*corresponding author: caflisch@bioc.uzh.ch

II. TESTS OF MARKOVIANITY

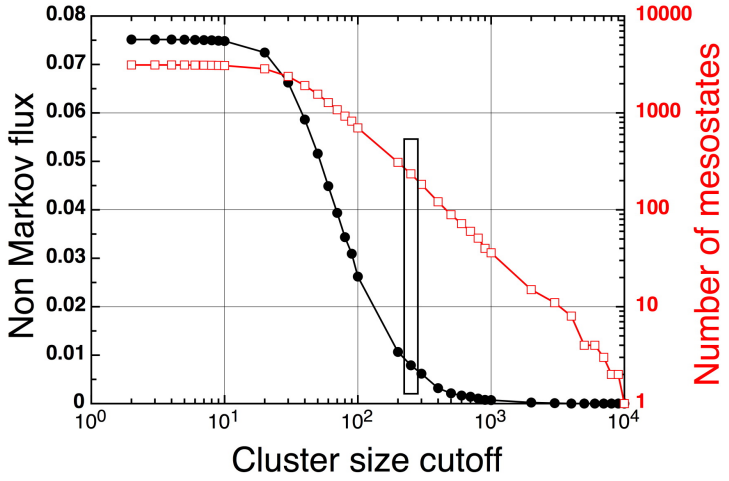

FIG. 2 Causal grouping of clusters. The causal grouped procedure, as explained in the main text of this paper, causally reassigns the conformers of the clusters whose size is less than a cutoff, to the clusters with size greater than the cutoff. The black curve in figure shows the values of the non-Markov flux on the new time series obtained with the causal grouping at a certain value of the cutoff. The curve has a sigmoidal shape with a midpoint corresponding to cluster size ~ 70 a flux 0.04. For cluster size ≥ 250 (rectangular box) the non-Markov flux is less than 0.01, which means that only about the 1% of the pathways in the new causal grouped time series are affected by long memory effects. There are 211 mesostates corresponding to a cluster size of 250.

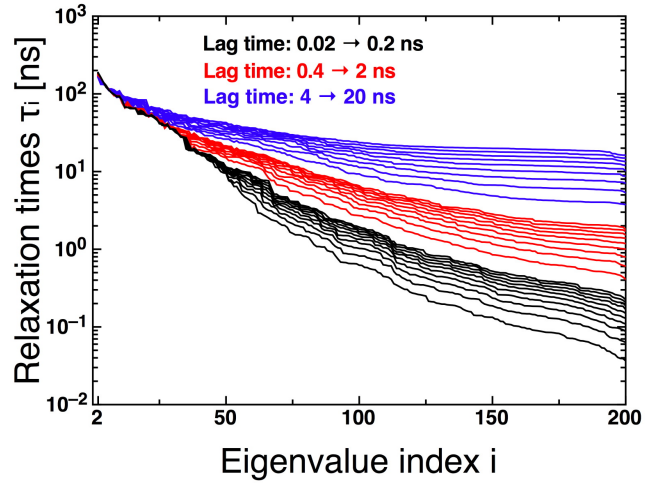

FIG. 3 Relaxation times of the decaying modes from the 200 causally-grouped mesostates. The slowest relaxations (i.e., indices 2-30) are robust with respect to changes in the lag time up to 20 ns. Note that a lag time of 20 ps was used for the Markov state model in the main text. To obtain the relaxation times from the transition matrices we calculated the reciprocal of the eigenvalues for the rate matrices $\mathbf{K}(\tau) = \mathbf{1} - \mathbf{T}(\tau)$, where $\mathbf{1}$ is the identity matrix.

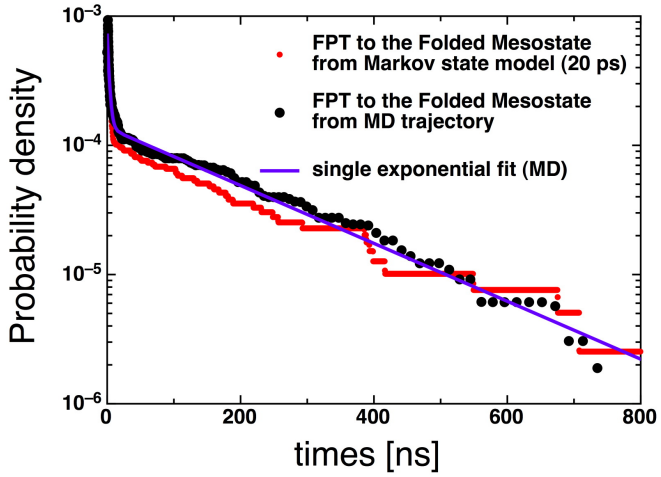

FIG. 4 Distribution of first passage time to the folded mesostate extracted directly from the MD trajectory, i.e., from the time series of causally grouped mesostates (black), and calculated by the Markov state model with a lag time of 20 ps (red). The solid line is a single exponential fit of the MD data. Note that folding is only slightly faster with the Markov state model than in the MD trajectory.

5 Estimation of protein folding probability from equilibrium simulations

(*Journal of Chemical Physics* (2005) **122**, 184901)

HTML ABSTRACT + LINKS

THE JOURNAL OF CHEMICAL PHYSICS **122**, 184901 (2005)

Estimation of protein folding probability from equilibrium simulations

Francesco Rao, Giovanni Settanni, Enrico Guarnera, and Amedeo Caflisch[a)]
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(Received 12 January 2005; accepted 23 February 2005; published online 6 May 2005)

The assumption that similar structures have similar folding probabilities (p_{fold}) leads naturally to a procedure to evaluate p_{fold} for every snapshot saved along an equilibrium folding-unfolding trajectory of a structured peptide or protein. The procedure utilizes a structurally homogeneous clustering and does not require any additional simulation. It can be used to detect multiple folding pathways as shown for a three-stranded antiparallel β-sheet peptide investigated by implicit solvent molecular dynamics simulations. © *2005 American Institute of Physics*. [DOI: 10.1063/1.1893753]

I. INTRODUCTION

The folding probability p_{fold} of a protein conformation saved along a Monte Carlo or molecular dynamics (MD) trajectory is the probability to fold before unfolding.[1] It is a useful measure of kinetic distance from the folded, i.e., functional state, and can be used to validate transition state ensemble (TSE) structures, which should have $p_{\mathrm{fold}} \approx 0.5$. Such validation consists of starting a large number of trajectories from putative TSE structures with varying initial distribution of velocities and counting the number of those that fold within a "commitment" time which has to be chosen much longer than the shortest time scales of conformational fluctuations and much shorter than the average folding time.[2] The concept of p_{fold} calculation originates from a method for determining transmission coefficients, starting from a known transition state[3] and the identification of simpler transition states in protein dynamics (e.g., tyrosine ring flips).[4] The approach has been used to identify the otherwise very elusive folding TSE by atomistic Monte Carlo off-lattice simulations of small proteins with a $G\bar{o}$ potential,[2,5] as well as implicit solvent MD (Refs. 6 and 7) and Monte Carlo[8] simulations with a physicochemical based potential. The number of trial simulations needed for the reliable evaluation of p_{fold} makes the estimation of the folding probability computationally very expensive. For this reason, here we propose a method to estimate folding probabilities for *all* structures visited in an equilibrium folding-unfolding trajectory without any additional simulation.

II. METHODS

A. Molecular dynamics simulations

Beta3s is a designed 20-residue sequence whose solution conformation has been investigated by NMR spectroscopy.[9] The NMR data indicate that beta3s in aqueous solution forms a monomeric (up to more than 1 mM concentration) triple-stranded antiparallel β sheet, in equilibrium with the denatured state.[9] We have previously shown that in implicit solvent[10] molecular dynamics simulations beta3s folds reversibly to the NMR solution conformation, irrespective of the starting structure.[11] Recently, four molecular dynamics simulations of beta3s were performed at 330 K for a total simulation time of 12.6 μs.[12] There are 72 folding events and 73 unfolding events and the average time required to go from the denatured state to the folded conformation is 83 ns. The 12.6 μs of simulation length is about two orders of magnitude longer than the average folding or unfolding time, which are similar because at 330 K the native and denatured states are almost equally populated.[12] For the p_{fold} analysis the first 0.65 μs of each of the four simulations were neglected so that along the 10 μs of simulations there are a total of 500 000 snapshots because coordinates were saved every 20 ps.

The simulations were performed with the program CHARMM.[13] Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field[13]). A mean field approximation based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute.[10] The two surface-tension-like parameters of the solvation model were optimized without using beta3s. The same force field and implicit solvent model have been used recently in molecular dynamics simulations of the early steps of ordered aggregation,[14] and folding of structured peptides,[10,11] as well as small proteins of about 60 residues.[15] Despite the absence of collisions with water molecules, in the simulations with implicit solvent the separation of time scales is comparable with that observed experimentally. Helices fold in about 1 ns,[16] β hairpins in about 10 ns,[16] and triple-stranded β sheets in about 100 ns,[12] while the experimental values are $\sim$0.1 μs,[17] $\sim$1 μs,[17] and $\sim$10 μs,[9] respectively.

B. Clusterization

The 500 000 conformations obtained from the simulations of beta3s (see above) were clustered by the leader algorithm.[18] Briefly, the first structure defines the first cluster and each subsequent structure is compared with the set of clusters found so far until the first similar structure is found. If the structural deviation (see below) from the first conformation of all of the known clusters exceeds a given thresh-

[a)]Author to whom correspondence should be addressed. FAX: +41 44 635 68 62. Electronic mail: caflisch@bioc.unizh.ch

 © 2005 American Institute of Physics

old, a new cluster is defined. The leader algorithm is very fast even when analyzing large sets of structures such as in the present work. The results presented here were obtained with a structural comparison based on the distance root mean square (DRMS) deviation considering all distances involving C_α and/or C_β atoms and a cutoff of 1.2 Å. This yielded 78 183 clusters. The DRMS and root mean square deviation of atomic coordinates (upon optimal superposition) have been shown to be highly correlated.[2] The DRMS cutoff of 1.2 Å was chosen on the basis of the distribution of the pairwise DRMS values in a subsample of the wild-type trajectories. The distribution shows two main peaks that originate from intracluster and intercluster distances, respectively (data not shown). The cutoff is located at the minimum between the two peaks. The main findings of this work are valid also for clusterization based on secondary structure similarity.[7,19]

C. Folding probability

For the computation of p_{fold} a criterion (Φ) is needed to determine when the system reaches the folded state. Given a clusterization of the structures, a natural choice for Φ is the visit of the most populated cluster which for structured peptides and proteins is not degenerate (other criteria are also possible, e.g., fraction of native contacts Q larger than a given threshold). Given Φ and a commitment time (τ_{commit}), the folding probability $p_{\text{fold}}(i)$ of a MD snapshot i is computed as[1,2]

$$p_{\text{fold}}(i) = \frac{n_f(i)}{n_t(i)}, \tag{1}$$

where $n_f(i)$ and $n_t(i)$ are the number of trials started from snapshot i which reach within a time τ_{commit} the folded state and the total number of trials, respectively.

Every simulation started from snapshot i can be considered as a Bernoulli trial of a random variable θ with value 1 (folding within τ_{commit}) or 0 (no folding within τ_{commit}). The variable θ has average and variance on the average of the form

$$\langle\theta\rangle = p_{\text{fold}} = \frac{1}{n_t}\sum_{i=1}^{n_t}\theta_i,$$
$$\sigma^2_{\langle\theta\rangle} = \frac{1}{n_t}p_{\text{fold}}(1-p_{\text{fold}}), \tag{2}$$

where n_t is the total number of trials and the accuracy on the p_{fold} value increases with n_t.

In Fig. 1 the distribution of the first passage time (fpt) to the folded state is shown. The double peak shape of the distribution provides evidence for the different time scales between *intra*basin and *inter*basin transitions. A value of 5 ns is chosen for τ_{commit} because events with smaller time scales correspond to the diffusion within the native free-energy basin, while events with larger time scales are transitions from other basins to the native one, i.e., folding/unfolding events.[12]

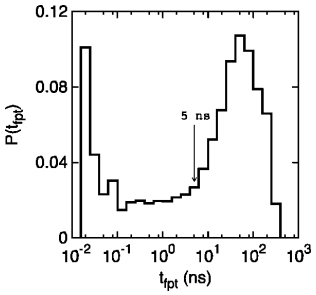

FIG. 1. Probability distribution for the first passage time (fpt) to the most populated cluster (*folded state*) of the DRMS 1.2 Å clusterization.

III. FOLDING PROBABILITY FROM EQUILIBRIUM TRAJECTORIES

The basic assumption of the present work is that conformations that are structurally similar have the same kinetic behavior, hence they have similar values of p_{fold}. Note that the opposite is not necessarily true as explained in Sec. IV for the TSE and the denatured state. To exploit this assumption, snapshots saved along a trajectory are grouped in structurally similar clusters.[20] Then the τ_{commit} segment of MD trajectory following each snapshot is analyzed to check if the folding condition Φ is met (i.e., the snapshot "folds"). For each cluster, the ratio between the snapshots which lead to folding and the total number of snapshots in the cluster is defined as the cluster $-p_{\text{fold}}$ (P_f^C; throughout the text uppercase P and lowercase p refer to folding probability for clusters and individual snapshots, respectively). This value is an approximation of the p_{fold} of any single structure in the cluster which is valid if the cluster consists of structurally similar conformations. In other words, the occurrence of the folding event for the snapshots of a given cluster can be considered as a Bernoulli trial of a random variable θ. The average of θ and variance on the average for the set of snapshots belonging to a given cluster α can be written as

$$P_f^C[\alpha] = \langle\theta\rangle = \frac{1}{W}\sum_{i=1}^{W}\theta_i, \quad i \in \alpha,$$
$$\sigma^2_{\langle\theta\rangle} = \frac{1}{W}P_f^C(1-P_f^C), \tag{3}$$

where W is the number of snapshots in cluster α. P_f^C is the average folding probability over a set of structurally homogeneous conformations. Using the clustering and the folding criterion Φ introduced above, values of P_f^C for the 78 183 clusters can be computed by Eq. (3), i.e., the number of conformations of the cluster that fold within 5 ns divided by the total number of conformations belonging to the cluster.

In this paper we provide evidence that the basic assumption mentioned above, that is, similar conformations have similar folding probabilities, holds in the case of beta3s, a three-stranded antiparallel β-sheet peptide investigated by MD.[12] Moreover, we show that the computationally expensive

TABLE I. DRMS clusters used for the calculation of P_f.

Cluster	P_f^C [a]	P_f [b]	$\sigma_{p_{\text{fold}}}$ [c]	N [d]	W [e]	W_{sample} [f]
1	0.00	0.03	0.04	150	144	15
2	0.11	0.05	0.06	150	449	15
3	0.06	0.05	0.07	120	36	12
4	0.08	0.07	0.08	140	555	14
5	0.10	0.08	0.06	100	10	10
6	0.13	0.12	0.18	160	911	16
7	0.25	0.16	0.07	80	4	4
8	0.23	0.20	0.31	150	141	15
9	0.21	0.22	0.15	140	178	14
10	0.12	0.23	0.20	120	48	12
11	0.57	0.25	0.14	140	14	14
12	0.05	0.27	0.19	100	19	10
13	0.23	0.29	0.38	140	391	14
14	0.08	0.30	0.15	120	12	12
15	0.72	0.35	0.23	130	129	13
16	0.19	0.38	0.18	130	26	13
17	0.38	0.44	0.39	160	16	16
18	0.38	0.51	0.28	160	16	16
19	0.65	0.60	0.29	100	20	10
20	0.57	0.61	0.35	70	7	7
21	0.48	0.63	0.32	140	27	14
22	0.74	0.65	0.40	140	539	14
23	0.68	0.66	0.18	140	28	14
24	0.38	0.71	0.24	130	13	13
25	0.50	0.72	0.20	100	2	2
26	0.82	0.76	0.31	170	17	17
27	0.50	0.78	0.14	120	12	12
28	0.78	0.78	0.22	180	18	18
29	0.70	0.79	0.19	130	189	13
30	0.77	0.79	0.17	150	30	15
31	0.85	0.81	0.11	130	13	13
32	0.91	0.83	0.20	140	401	14
33	0.90	0.85	0.27	100	20	10
34	0.85	0.85	0.10	120	48	12
35	0.94	0.88	0.13	170	1990	17
36	0.71	0.94	0.07	70	7	7
37	0.95	0.95	0.06	150	855	15

[a]Cluster $-p_{\text{fold}}$ [P_f^C, Eq. (3)].
[b]Traditional, i. e., computationally expensive P_f value [Eq. (4)].
[c]Standard deviation of p_{fold} in a cluster [Eq. (5)].
[d]Total number of trials used to evaluate P_f. For every structure $n_t=10$ trials were performed ($N=n_t\,W_{\text{sample}}$) except for clusters 7 and 25 for which 20 and 50 trials were performed, respectively.
[e]Number of snapshots in the cluster.
[f]Number of snapshots used to evaluate P_f. The W_{sample} subset was obtained by selecting structures in a cluster every $\lfloor W/W_{\text{sample}} \rfloor$ saved conformations.

$$P_f[\alpha] = \frac{1}{W}\sum_{i=1}^{W} p_{\text{fold}}(i), \quad i \in \alpha, \tag{4}$$

which is measured by starting several simulations from each snapshot i in the cluster α with W snapshots, is well approximated by P_f^C whose evaluation is straightforward.

To test the assumption that similar structures have similar p_{fold} and to compare the values of P_f^C with those obtained from the standard approach,[1] folding probabilities P_f were computed for the structures of 37 clusters by starting several 5 ns MD runs from each structure and counting those that fold [Eqs. (1) and (4)]. The 37 clusters chosen among the 78 183 include both high- and low-populated clusters with P_f^C values evenly distributed in the range between 0 and 1 (see Table I). In the case of large clusters a subset of snapshots is considered for the computation of P_f. In those cases W is replaced in Eq. (4) by $W_{\text{sample}} < W$ that is the number of snapshots involved in the calculation.

The standard deviation of p_{fold} in a cluster is computed as

$$\sigma_{p_{\text{fold}}} = \sqrt{\langle (p_{\text{fold}}(i) - P_f[\alpha])^2 \rangle_{i\in\alpha}}. \tag{5}$$

In the case of full kinetic inhomogeneity, i.e., random grouping of snapshots, the p_{fold} value for all snapshots in a given cluster will be equal to 0 or 1, indicating the coexistence (in the same cluster) of structures that either exclusively fold or unfold. In this case $\sigma_{p_{\text{fold}}}$ reflects the Bernoulli distribution.[19] Figure 2(a) shows that, even when only $n_t=10$ runs per snapshot are used to compute p_{fold}, $\sigma_{p_{\text{fold}}}$ values are not compatible with those of a Bernoulli distribution. Moreover the values of the standard deviation decrease when the number of trials n_t increases, as reported in Fig. 2(b) for two sample clusters. The asymptotic value of $\sigma_{p_{\text{fold}}}(n_t \to \infty)$ for these two data sets is of 0.05 and 0.2. This value cannot reach zero because snapshots in a cluster are similar but not identical. These results suggest that snapshots inside the same cluster are kinetically homogeneous and a statistical description of p_{fold} can be adopted, that is, folding probabilities are computed as cluster averages (instead of single snapshots) by means of P_f and P_f^C.

We still have to verify that P_f^C indeed approximates the computationally expensive P_f. Namely, for the 37 clusters mentioned above a correlation of 0.89 between P_f^C and P_f is found with a slope of 0.86 (see Fig. 3(a) and Table I), indicating that the procedure is able to estimate folding probabilities for clusters on the folding-transition barrier ($P_f \sim 0.5$) as well as in the folding ($P_f \sim 1.0$) or unfolding ($P_f \sim 0.0$) regions. The error bars for P_f^C in Fig. 3(a) are derived from the definition of variance given in Eq. (3). In the same spirit of Eq. (3) the folding probability P_f and its variance are written as

$$P_f = \langle\theta\rangle = \frac{1}{N}\sum_{i=1}^{N} \theta_i, \qquad \sigma^2_{\langle\theta\rangle} = \frac{1}{N} P_f(1-P_f), \tag{6}$$

where $N=\Sigma n_t$ is the total number of runs and θ is equal to 1 or 0, if the run folded or unfolded, respectively. Note that the same number of runs n_t has been used for every snapshot of a cluster. The large vertical error bars in Fig. 3(a) correspond to clusters with less than ten snapshots. The largest deviations between P_f and P_f^C are around the 0.5 region. This is due to the limited number of crossings of the folding barrier observed in the MD simulation [Fig. 3(b), around 70 events of folding[12]]. Improvements in the accuracy for the estimation of P_f are achieved as the number of folding events, i.e., the simulation time, increases [Figs. 3(c)–3(e)].

The two main results of this study, i.e., the kinetic homogeneity of the clusters and the validity of P_f^C as an ap-

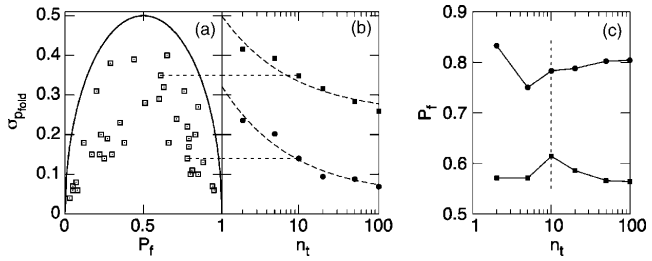

FIG. 2. Standard deviation $\sigma_{p_{\text{fold}}}=\sqrt{\langle(p_{\text{fold}}(i)-P_f[\alpha])^2\rangle_{i\in\alpha}}$ of the p_{fold} for the 37 DRMS clusters used in the study. (a) $\sigma_{p_{\text{fold}}}$ as a function of P_f compared to a Bernoulli distribution (solid line). Ten trials were performed for each snapshot. The largest values for the standard deviation are located around the 0.5 region and this is probably due to the Bernoulli process ($\theta=0,1$) used for the calculation of p_{fold}. (b) $\sigma_{p_{\text{fold}}}$ dependence on the number of trials used to evaluate p_{fold}. The dashed curves are fits with a $(a/\sqrt{x})+b$ function. The horizontal dashed lines are drawn to help identifying in (a) the two clusters used in (b). (c) Dependence of P_f on the number of trials n_t for the two clusters used in (b).

proximation of P_f, are robust with respect to the choice of the clusterization. Similar results can be obtained also with different flavors of conformation space partitioning, as long as they group together structurally homogeneous conformations, e.g., clusterization based on root mean square deviation of atomic coordinates (RMSD) or secondary structure strings.[19] The latter are appropriate for structured peptides but not for proteins with irregular secondary structure because of string degeneracy. Note that partitions based on order parameters (like native contacts) are usually unsatisfactory and not robust. This is mainly due to the fact that clusters defined in this way are characterized by large structural heterogeneities.[7]

IV. ANALYSIS OF TRANSITION STATE ENSEMBLE

The folding probability of structure i is estimated as $p_{\text{fold}}(i)=P_f^C[\alpha]$ for $i\in\alpha$. This approximation allows to plot the pairwise RMSD distribution of beta3s structures with $p_{\text{fold}}>0.51$ (native state), $0.49<p_{\text{fold}}<0.51$ (transition state ensemble, TSE), and $p_{\text{fold}}<0.49$ (denatured state) [Fig. 4(a)]. For the native state, the distribution is peaked around low values of RMSD (~1.5 Å) indicating that structures with $p_{\text{fold}}>0.51$ are structurally similar and belong to a nondegenerate state. The statistical weight of this group of structures is 49.4% and corresponds to the expected statistics for the native state because the simulations are performed at the melting temperature. In the case of TSE, the distribution is broad because of the coexistence of heterogeneous structures. This scenario is compatible with the presence of multiple folding pathways. Beta3s folding was already shown to involve two main average pathways depending on the sequence of formation of the two hairpins.[7,11] Here, a *naive* approach based on the number of native contacts[11] is used to structurally characterize the folding barrier. TSE structures

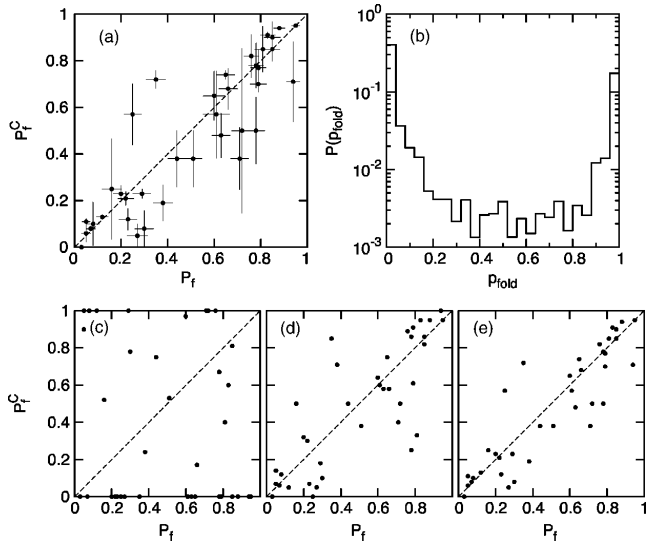

FIG. 3. Cluster folding probability P_f^C. (a) Scatter plot of P_f^C vs P_f. The DRMS 1.2 Å clusterization and the folding criterion Φ (reaching the most populated cluster within $\tau_{\text{commit}}=5$ ns) were used. (b) Probability distribution of the p_{fold} value for the 500 000 snapshots saved along the 10 μs MD trajectory. The folding probability for snapshot i is computed as $p_{\text{fold}}(i)=P_f^C[\alpha]$ for $i\in\alpha$. (c–e) Scatter plot of P_f^C vs P_f for 1.0, 5.0, and 10 μs of simulation time, respectively.

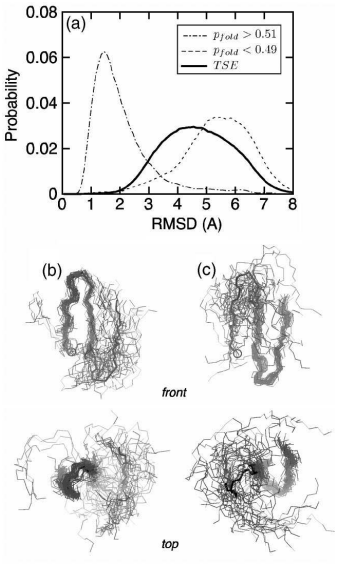

FIG. 4. Transition state ensemble (TSE) of beta3s. (a) RMSD pairwise distribution for structures with $p_{\text{fold}} > 0.51$ (native state), $0.49 < p_{\text{fold}} < 0.51$ (TSE), and $p_{\text{fold}} < 0.49$ (denatured state). (b) Type I and (c) type II transition states (thin lines). Structures are superimposed on residues 2–11 and 10–19 with an average pairwise RMSD of 0.81 and 0.82 Å for type I and type II, respectively. For comparison, the native state is shown as a thick line with a circle to label the *N* terminus.

with number of native contacts of the first hairpin greater than the ones of the second hairpin are called type I conformations [Fig. 4(b)], otherwise they are called type II [Fig. 4(c)]. In both cases the transition state is characterized by the presence of one of the two native hairpins formed while the rest of the peptide is mainly unstructured. These findings are also in agreement with the complex network analysis of beta3s reported in Ref. 7. Finally, the denatured state shows a broad pairwise RMSD distribution around even larger values of RMSD (~5.5 Å), indicating the presence of highly heterogeneous conformations.

V. CONCLUSIONS

Two main results have emerged from the present study. First, snapshots grouped in structurally homogeneous clusters are characterized by similar values of p_{fold}. This result justifies the use of a statistical approach for the study of the kinetic properties of the structures sampled along a simulation. Second, given a set of structurally homogeneous clusters and a folding criterion, it is possible to obtain a first approximation of the folding probability for every structure sampled along an equilibrium folding-unfolding simulation. Thus, the cluster $-p_{\text{fold}}$ is a quantitative measure of the kinetic distance from the native state and is computationally very cheap.[21] Furthermore, it can be used to detect multiple folding pathways. The accuracy in the identification of the transition state ensemble improves as the number of folding events observed in the simulation increases. Recently the cluster p_{fold} approach has been used to identify the transition state ensemble of a large set of beta3s mutants (for a total of 0.65 ms of simulation time[22]), which would have been impossible with traditional methods. As a further application, the cluster $-p_{\text{fold}}$ procedure can be used to validate TSE conformations obtained by wide-spread $G\bar{o}$ models.

ACKNOWLEDGMENTS

The authors thank Stefanie Muff for useful and stimulating discussions and comments to the manuscript. They also thank Dr. Emanuele Paci for interesting discussions. The molecular dynamics simulations were performed on the Matterhorn Beowulf cluster at the Informatikdienste of the University of Zurich. They thank C. Bollinger, Dr. T. Steenbock, and Dr. A. Godknecht for setting up and maintaining the cluster. This work was supported by the Swiss National Science Foundation under Grant No. 205321-105946/1.

[1] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. I. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[2] I. Hubner, J. Shimada, and E. Shakhnovich, J. Mol. Biol. **336**, 745 (2004).
[3] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).
[4] S. H. Northrup, M. R. Pear, C. Y. Lee, J. A. McCammon, and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **79**, 4035 (1982).
[5] L. Li and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **98**, 13014 (2001).
[6] J. Gsponer and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **99**, 6719 (2002).
[7] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[8] P. Lenz, B. Zagrovic, J. Shapiro, and V. S. Pande, J. Chem. Phys. **120**, 6769 (2004).
[9] E. De Alba, J. Santoro, M. Rico, and M. A. Jiménez, Protein Sci. **8**, 854 (1999).
[10] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Struct., Funct., Genet. **46**, 24 (2002).
[11] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **97**, 10780 (2000).
[12] A. Cavalli, U. Haberthür, E. Paci, and A. Caflisch, Protein Sci. **12**, 1801 (2003).
[13] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[14] J. Gsponer, U. Haberthür, and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **100**, 5154 (2003).
[15] J. Gsponer and A. Caflisch, J. Mol. Biol. **309**, 285 (2001).
[16] P. Ferrara, J. Apostolakis, and A. Caflisch, J. Phys. Chem. B **104**, 5000 (2000).
[17] W. A. Eaton, V. Munoz, J. Hagen, S G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter, Annu. Rev. Biophys. Biomol. Struct. **29**, 327 (2000).
[18] J. A. Hartigan, *Clustering Algorithms* (Wiley, New York, 1975).
[19] See EPAPS Document No. E-JCPSA6-122-505519 for Supplementary Material. This document can be reached via a direct link in the HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).
[20] Making a structural cluster analysis is equivalent to a partition of the conformation space. Given an appropriate partition, it is possible to analyze the dynamic behavior in terms of symbol sequences generated by the simulation. The symbol sequences describe the time evolution of the trajectories in a coarse-grained way. The mapping from the conformation space to the symbol space is called *symbolic dynamics* [C. Beck and F. Schloegl, *Thermodynamics of Chaotic Systems* (Cambridge University Press, Cambridge, 1993)]. Subsets α_i, also called *cells* or *clusters*, of different size and shape are used to partition the conformation space. The subsets are disjoint and cover the entire conformation space.
[21] The computation of P_f^C presented in this work takes few seconds on a desktop computer.
[22] G. Settanni, F. Rao, and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **102**, 628 (2005).

6 Pathways and intermediates of amyloid fibril formation

(*Journal of Molecular Biology* (2007), **379**, 917-924)

Article

doi:10.1016/j.jmb.2007.09.090 *J. Mol. Biol.* (2007) **374**, 917–924

JMB

Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

COMMUNICATION

Pathways and Intermediates of Amyloid Fibril Formation

Riccardo Pellarin, Enrico Guarnera and Amedeo Caflisch*

Department of Biochemistry
University of Zürich
Winterthurerstrasse 190
CH-8057 Zürich, Switzerland

Received 18 July 2007;
received in revised form
13 September 2007;
accepted 28 September 2007
Available online
4 October 2007

The lack of understanding of amyloid fibril formation at the molecular level is a major obstacle in devising strategies to interfere with the pathologies linked to peptide or protein aggregation. In particular, little is known on the role of intermediates and fibril elongation pathways as well as their dependence on the intrinsic tendency of a polypeptide chain to self-assembly by β-sheet formation (β-aggregation propensity). Here, coarse-grained simulations of an amphipathic polypeptide show that a decrease in the β-aggregation propensity results in a larger heterogeneity of elongation pathways, despite the essentially identical structure of the final fibril. Protofibrillar intermediates that are thinner, shorter and less structured than the final fibril accumulate along some of these pathways. Moreover, the templated formation of an additional protofilament on the lateral surface of a protofibril is sometimes observed as a collective transition. Conversely, for a polypeptide model with a high β-aggregation propensity, elongation proceeds without protofibrillar intermediates. Therefore, changes in intrinsic β-aggregation propensity modulate the relative accessibility of parallel routes of aggregation.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: amyloid protofibrils; fibril growth; aggregation pathways; molecular dynamics simulations; Alzheimer's disease

Edited by F. E. Cohen

The link between protein aggregates and progressive neurodegenerative pathologies, like Alzheimer's, Parkinson's, Huntington's and prion diseases, exists but is not clear.[1,2] Despite the medical relevance of these devastating diseases, little is known about the aggregation process itself and, most importantly, how to safely inhibit the formation of toxic species. Experimental evidence indicates that early aggregates, e.g. soluble oligomers and protofibrils, have a critical role in promoting pathological effects in amyloid disorders.[3,4] As an example, the E22G mutation of the Alzheimer's peptide (Aβ) enhances protofibril formation,[5] and plaque formation is more aggressive than for wild-type Aβ in transgenic mice.[6] Also, mutations of α-synuclein that are related to early-onset forms of Parkinson's disease can produce protofibrils efficiently.[7] Yet, the molecular details and the mechanisms leading to the toxicity of these prefibrillar aggregates are only partially understood. In fact, the transient character of oligomeric precursors hinders the complete understanding of their formation process and structural details.

*Corresponding author. E-mail address: caflisch@bioc.uzh.ch.

The available experimental evidence *in vitro* indicates that the kinetics of fibril formation are complex and can be often separated into a nucleation (or lag) phase and an elongation phase,[8] followed by the equilibrium between isolated polypeptides and the fibrils.[9] Multistep kinetics with the presence of intermediates have also been reported.[10] Pathways of fibril formation, fibril morphologies and stability of protofibrillar intermediates are influenced strongly by experimental conditions (e.g. protein concentration, pH and ionic strength),[11] and elongation rates can depend on the stability of aggregation prone folding intermediates.[12]

Theoretical models have been developed to investigate the amyloid aggregation mechanism[13–15] and predict the rates[16] but strong assumptions like the irreversible association of polypeptide chains onto the fibril[13,16] are not consistent with the interpretation of experimental results.[9,17] Computer simulations using low-resolution models, which employ a simplified representation of protein geometry and energetics, have provided insights into the basic physical principles underlying protein aggregation in general,[18–20] and ordered amyloid aggregation.[21–28] However, they do not explain the wide range of aggregation processes emerging from a variety of

0022-2836/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.

biophysical studies.[11,29] Atomistic models have shed some light on oligomeric aggregates and the very early steps of fibril formation,[30–36] but all-atoms simulations aimed at reproducing the kinetics and investigating the pathways of fibril formation are computationally expensive and difficult to analyze.

Earlier, we developed a phenomenological coarse-grained model of an amphipathic polypeptide and used it for exploring the kinetics of nucleation and the rates of and elongation by Langevin dynamics simulations.[37] To allow for efficient sampling, the conformational landscape of the isolated monomer was simplified such that only two states are considered: the amyloid-competent (β) and the amyloid-protected (π) states (Figure 1). In the β-state, the parallel orientation of the two intramolecular dipoles favors ordered aggregates with intermolecular dipolar interactions parallel with the fibril axis. Conversely, the π-state represents the ensemble of all polypeptide conformations that are not compatible with self-assembly into a fibril. At physiological temperature the isolated monomer undergoes a reversible isomerization from the π-state to the β-state. The energy difference between these two states can be interpreted as the β-aggregation propensity of a polypeptide sequence. For instance when $dE = E_\pi - E_\beta = 0.0$ kcal/mol, the π and β states are equally populated, whereas for $dE = -1.5$ kcal/mol and -2.5 kcal/mol the π-state is about 15 and 100 times more populated than the β state, respectively. It was found that despite the essentially identical structure of the final fibril, ordered aggregation of a polypeptide with a stable β-state follows a pathway devoid of stable intermediates, while on-pathway micellar oligomers (with hydrophilic surface and hydrophobic interior) were observed during the nucleation phase of a polypeptide with a β-state that is marginally stable. In other words, high and low β-prone sequences show significantly different nucleation processes. These two models are termed β-stable and β-unstable, respectively, and the passage from one regime to the other was achieved by varying solely the parameter dE.[37] The focus of our previous study was on the nucleation phase, while the elongation mechanism and pathway(s) were not investigated. Here, for each of four

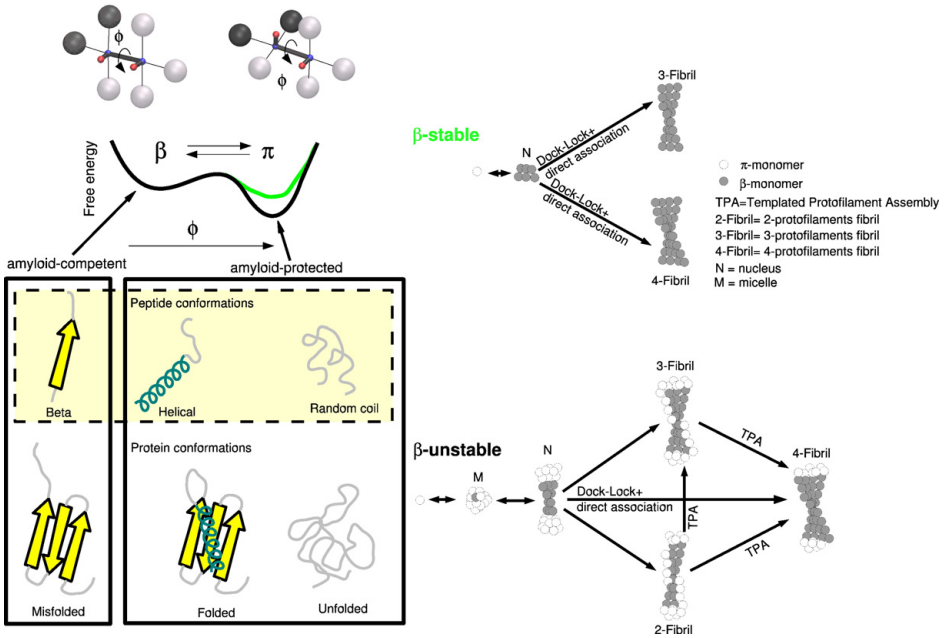

Figure 1. The model and aggregation pathways. Left: Sticks and beads representations of the monomer in the amyloid-competent state β and the amyloid-protected state π. The large spheres are hydrophobic (black) and hydrophilic (gray), while the two dipoles are shown with small red and blue spheres. The β and π states of the monomer are shown on top of the two corresponding minima of the free energy, plotted as a function of the dihedral angle ϕ of the two dipoles. Note that the population of monomers in the β-state decreases by lowering the free energy of the π-state, as indicated by the green and black profiles. For each value of the β-aggregation propensity dE ($dE = E_\pi - E_\beta = -1.5$, -2.0, -2.25, -2.5 kcal/mol) 100 Langevin dynamics runs with different initial assignments of the velocities were started from 125 monomers uniformly distributed in a box with random orientations. All simulations were carried out at a temperature of 310 K and a concentration of 8.5 mM with the same force-field parameters as those used previously.[37] Results discussed in this work refer mainly to the β-stable ($dE = -1.5$ kcal/mol) and the β-unstable ($dE = -2.5$ kcal/mol) models. Right: Observed aggregation pathways for the β-stable and β-unstable models. The elongation pathways of the latter are more heterogeneous than those of the former.

polypeptide models (four values of dE that range from β-stable to β-unstable) 100 Langevin dynamics runs were performed to explore the elongation phase; i.e. the pathway(s) leading from the nucleus to the final fibril.

The present work was motivated by the following two questions: what is the influence of the intrinsic β-aggregation propensity on the mechanism of fibril elongation? and are there multiple pathways and/or intermediates? From a detailed analysis of the simulations (started from 125 coarse-grained monomers in a monodisperse state), a rich scenario of alternative pathways, some with prefibrillar intermediates, emerges only for monomers with a low β-aggregation propensity. The simulation results go beyond the fibril formation mechanisms suggested on the basis of biophysical measurements, and have strong implications for the design of inhibitors of amyloid aggregation.

Terminology

A rigorous terminology for the early aggregates and intermediates of amyloid self-assembly observed *in vitro* has been recently summarized.[38,39] Because the computer simulations allow for the detailed investigation of individual oligomers as well as prefibrillar states and the final fibril, it is useful and straightforward to define the following nomenclature: a protofilament is a file of monomers with intermolecular dipolar interactions parallel with its axis; a protofibril is a transient structure that consists of two to three protofilaments with large unstructured regions; and the final fibril is a fully ordered aggregate of three to four protofilaments. In the model used here, the fibril is stabilized by intermolecular dipolar interactions within each protofilament and van der Waals interactions between hydrophobic beads.[37]

Aggregation state network

An aggregate consists of monomers whose mutual minimal distances are less than 6 Å, and it is isolated using a clustering procedure as described.[37] Three progress variables are used to monitor the aggregation process: the size of the largest aggregate N_{la}, the number of monomers in the β-state within the largest aggregate N_{la}^{β}, and the number of protofilaments in the largest aggregate N_{la}^{pf}. Note that the range of N_{la} is limited by the size of the simulated system ($1 \leq N_{la} \leq 125$). The number of protofilaments within a single aggregate is calculated by counting the files of monomers in the β-state with intermolecular dipolar interactions. Let N_f be the number of such files present into a given aggregate, and $\omega_1, \cdots, \omega_{Nf}$ the number of monomers in each file (with $\omega_i > 10$ to reduce noise). The number of protofilaments in aggregate a, N_a^{pf}, is thus defined as:

$$N_a^{pf} = \frac{\left(\sum_{i=1}^{N_f} \omega_i\right)^2}{\sum_{i=1}^{N_f} \omega_i^2} \quad (1)$$

This definition prevents counting small isolated files whose formation is a result of thermal fluctuations, enhancing the signal to noise ratio with respect to N_f. Two limiting cases are useful to explain this variable. In the case that all files have the same size (i.e. $\omega_1 = \ldots = \omega_{Nf}$), the protofilament number N_a^{pf} is equal to the number of files N_f. In the case where a single ω_j predominates ($\omega_i \gg \omega_k$ for all k different from i) N_a^{pf} tends to 1. The number of protofilaments in the largest aggregate N_{la}^{pf} is thus the function N_a^{pf} applied to the largest of all aggregates present in the simulation volume. Selected time series of N_{la}, N_{la}^{β} and N_{la}^{pf} are reported in Figure 2.

The aggregation state network (Figure 3) is a graph in which states and direct transitions observed during the Langevin dynamics simulations are displayed as nodes and links, respectively.[40] Furthermore, the size of each node reflects the statistical weight of the corresponding state. In this way, metastable states and their dynamic connectivity are illustrated without requiring projections onto arbitrarily chosen reaction coordinates.[41] Micellar oligomers (white nodes, $N_{la} \sim 20, N_{la}^{pf} = 0$), which are spherical aggregates whose core consists of the hydrophobic spheres of the monomers (see inset A of Figure 3),[37] and fibrils (red nodes, $N_{la} \sim 100$, $N_{la}^{pf} = 4$) are the most populated states during the lag phase and the final equilibrium, respectively. Strikingly, a greater variety of aggregation mechanisms emerges for the β-unstable (Figure 3, bottom) than the β-stable polypeptide model (Figure 3, top). In particular, the former shows the presence of intermediates, i.e. protofibrils consisting of only two (green nodes) or three (blue nodes) protofilaments. Moreover, the aggregation state network qualitatively illustrates that the protofibrils are metastable and it displays broad transition regions between the two-protofilament state and the three-protofilament state, as well as between the latter and the final fibril.

Templated protofilament assembly

Previously, the elongation rate was found to increase according to the population of the amyloid-competent state,[37] but the underlying mechanism of elongation was not investigated. Using the Markov chain formalism (see the Supplementary Data) it is possible to estimate the rate of association of a monomer to a fibril followed by the isomerization from the amyloid-protected state to the amyloid-competent state (k_{fibril}). An alternative process is the monomer isomerization in the solvent followed by association ($k_{solvent}$). In their analytical model of fibril elongation, Massi and Straub have illustrated these two pathways as the route of monomer association to the fibril followed by isomerization (deposition and reorganization in Figure 4 of Massi & Straub[14]) and the route of direct association (direct deposition in Figure 5 of Massi & Straub[14]). The former pathway corresponds to the dock-lock mechanism.[42–44] Hence, the ratio $k_{fibril}/k_{solvent}$ measures the efficiency of the dock-lock mechanism; it is

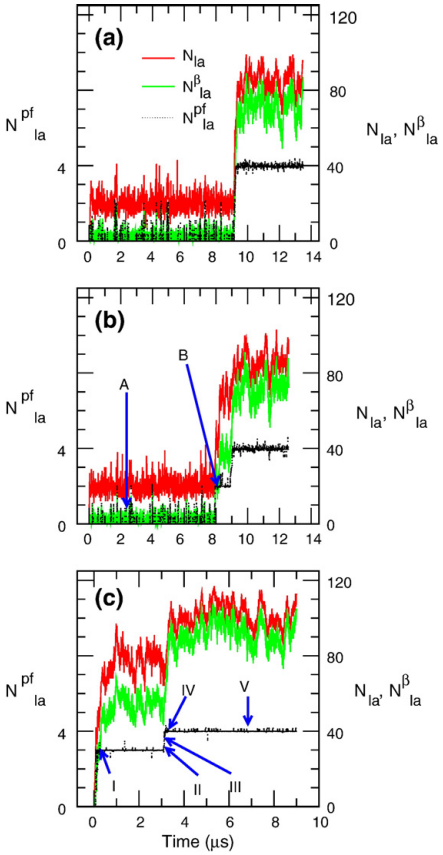

Figure 2. Protofibrillar intermediates and pathway heterogeneity. The time-series of three progress variables are used to monitor the evolution of the largest aggregate (1a) in the β-unstable simulations: The number of protofilaments N_{la}^{pf} (black curve with the *y*-axis description on the left; note that this quantity is evaluated by equation (1) and can be non-integer), the size of the largest aggregate N_{la} and the number of monomers in β-state N_{la}^{β} (red and green curves, respectively, with the *y*-axis description on the right). The three runs shown are representative of (a) elongation without intermediates, and (b) with two-filament or (c) three-filament protofibrillar intermediates. Templated protofilament assembly is observed at about 9 μs in (b) and at about 3 μs in (c), and the snapshots labeled are shown in Figure 3.

3.6 for the β-unstable model and 6.6 for the β-stable model. In both cases, the rate of conversion of a monomer bound to a fibril exceeds that in solution, suggesting that the elongation is dominated by a dock-lock mechanism. Nevertheless, this mechanism does not exclude collective conversions.

Representative time series of the number of protofilaments N_{la}^{pf} are shown by a black curve in Figure 2 for the β-unstable model. Metastable intermediates are observed in about half of the runs (see Supplementary Data). Interestingly, during some of the fast transitions from a three-protofilament aggregate to the final fibril (or sporadically from two to three-protofilament protofibrils) the size of the largest aggregate (red line) does not change significantly, whereas its number of monomers in the β-state (green line) increases abruptly, e.g. at about 9 μs and 3 μs in Figure 2(b) and (c), respectively. The collective conversion of monomers from the amyloid-protected to the amyloid-competent state is a consequence of the templated assembly of the fourth filament on the metastable protofibril consisting of three protofilaments (Figure 3 insets I–V). In other words, a file of monomers in the amyloid-protected conformation accumulates, first without forming intermolecular dipolar interactions, along the exposed hydrophobic surface of the three-protofilament aggregate (blue monomers in inset I). This event is then followed by a collective transition during which all monomers in the file convert to the β-state, which is stabilized by both intermolecular dipole interactions within the fourth protofilament and van der Waals interactions with monomers in the other three protofilaments (insets II–IV). The templated-assembly mechanism observed in the simulations is consistent with measurements of insulin aggregation by atomic force microscopy.[45] Moreover, protofibril maturation into fibrils is irreversible under the conditions used in the present simulations, i.e. 310 K and 8.5 mM (see Figure 2). Irreversibility has been suggested on the basis of the temporal increase in average protofibril size measured by quasi-elastic light-scattering spectroscopy.[46]

Analysis of the time series of the β-stable model does not reveal any event of templated protofilament formation. In fact, fibrils composed of three protofilaments contain as many monomers in the β-state as the mature four-protofilament fibril (see Figure 4(d)); thus, the formation of the fourth protofilament corresponds to a redistribution of monomers in the β-state among the protofilaments.

Size and structural characterization of protofibrils

The size distribution of the two and three-protofilament aggregates are different and depend on the β-aggregation propensity of the monomer (Figure 4). During the elongation phase, intermediates with two protofilaments are observed mainly for the β-unstable model (peak at $N_{la} \sim 70$). By raising the β-aggregation propensity (from $dE = -2.5$ kcal/mol to $dE = -1.5$ kcal/mol) there is a decrease in the average aggregation size of two-protofilament aggregates. Protofibrils consisting of three protofilaments are observed during the elongation phase of all models. Notably, by increasing the β-aggregation tendency, the number of runs with on-pathway intermediates decreases monotonically, which reflects the lower heterogeneity of pathways for the β-stable model.

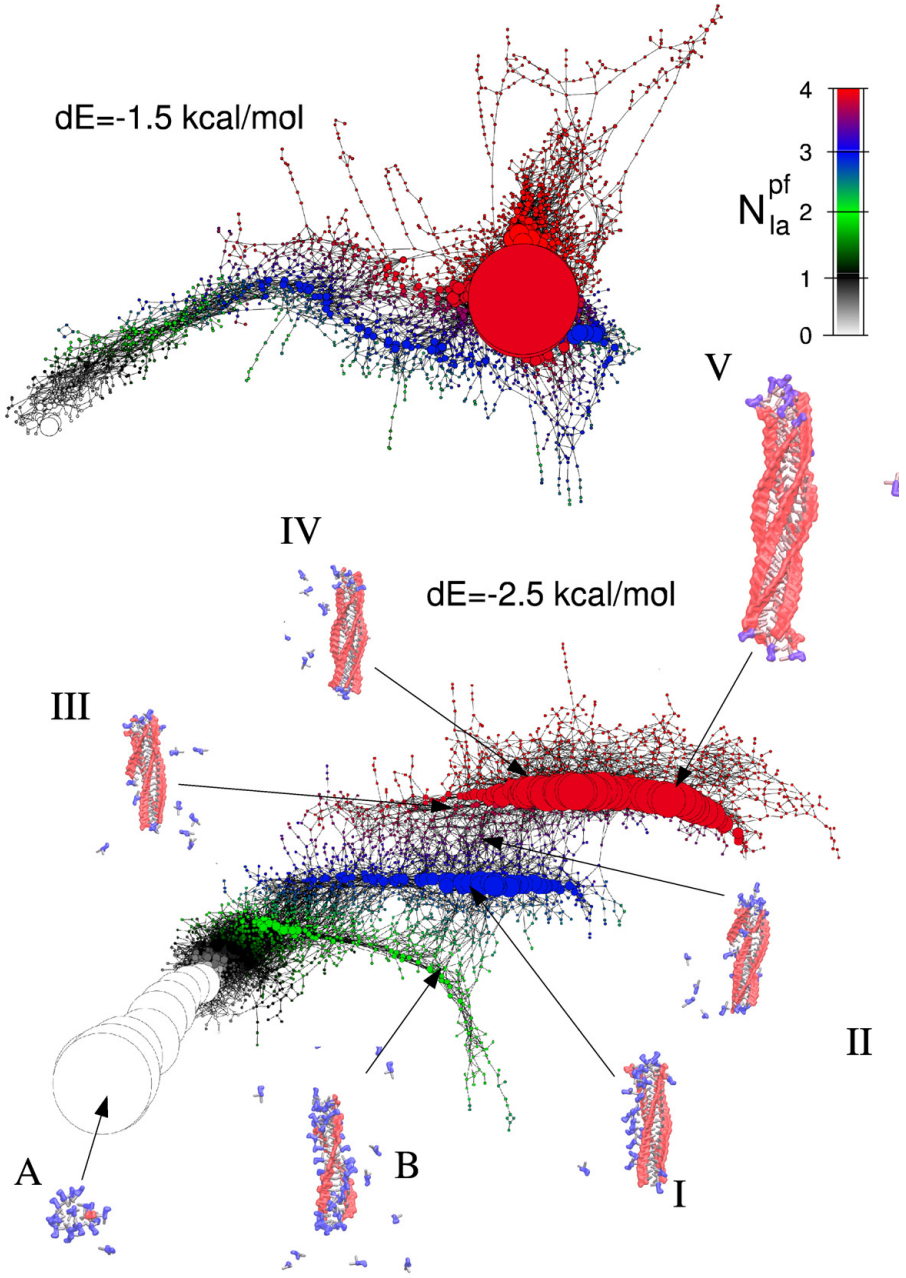

Figure 3. Aggregation state network. The size of the largest aggregate N_{la} and its number of protofilaments N_{la}^{pf} were used to cluster all snapshots into states (i.e. nodes of the network). The size and color of the nodes correspond to the statistical weight and the number of protofilaments N_{la}^{pf}, respectively. Links are direct transitions within 0.5 ns (10,000 steps of 50 fs each) of Langevin dynamics. All the states and the transitions that have been explored by the simulations are represented in these networks. Note the much higher heterogeneity of protofibrillar intermediates for the β-unstable ($dE = -2.5$ kcal/mol, bottom) than the β-stable ($dE = -1.5$ kcal/mol, top) model. The insets show the structures of the largest aggregates from the snapshots labeled in Figure 2. In these structures, monomers in the amyloid-competent conformer β and amyloid-protected conformer π are in red and blue, respectively. Furthermore, hydrophobic spheres are gray and hydrophilic spheres are not shown for visual clarity.

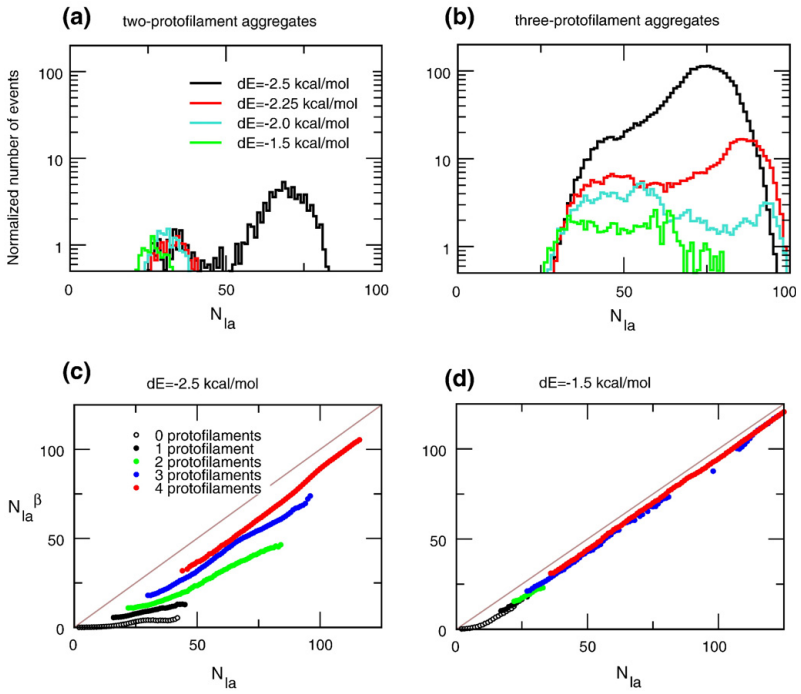

Figure 4. Size distribution of (a) two-protofilament and (b) three-protofilament protofibrils during fibril growth. The histograms are built by counting the trajectory frames in which the largest aggregate contains either two or three protofilaments. The frames are collected only during the elongation phase, i.e. after the nucleation step and before reaching the final monomer/fibril equilibrium. Average value of the number of monomers in β-state contained into the largest aggregate, as a function of the size of the largest aggregate for the (c) β-unstable and (d) the β-stable models.

For the β-unstable model protofibrils are thinner, shorter and more disordered than the final fibril. The protofibrils and fibrils of this model often present deposits of monomers in the π-state that are not involved in intermolecular dipole interactions and are highly disordered (blue monomers in the insets of Figure 3). The ratio between the number of monomers in the β-state and the total number of monomers N_{la}^{β}/N_{la} is significantly smaller than 1, even for fibrils consisting of four protofilaments (Figure 4(c)). The deviation is due mainly to the fibril ends that are populated by monomers in the π-state (see Figure 3 inset V). Furthermore, protofibrils with two or three protofilaments contain less monomers in the β-state than the four-protofilament fibril of the same size. Conversely, for the β-stable model the N_{la}^{β}/N_{la} ratio is always close to 1, and aggregates of three protofilaments can have more than 100 monomers (Figure 4(d)).

Conclusions

The self-assembly process of an amphipathic polypeptide has been investigated by multiple Langevin dynamics simulations using a coarse-grained model whose simplicity allows for the sampling of hundreds of fibril formation events. By varying a single parameter of the model, namely the relative stability of the amyloid-competent and amyloid-protected states of the polypeptide (β-aggregation propensity), interesting insights into elongation pathways and protofibrillar intermediates have been obtained. Two main observations emerge from the simulation results.

First, the roughness of the free-energy surface governing the aggregation process and the heterogeneity of pathways of fibril elongation increase by reducing the β-aggregation propensity. Hence, a mutation that decreases the β-aggregation tendency could result in greater variety of prefibrillar aggregates. Interestingly, these simulation results provide a possible explanation for the enhanced *in vitro* formation of oligomers and protofibrils of the Arctic mutant (E22G) of the Alzheimer's Aβ peptide,[5] and the A30P mutant of α-synuclein.[7] In fact, among the 20 standard amino acids, glycine and proline residues have the weakest propensity of β-sheet formation,[47] and β-aggregation.[48]

Second, a mechanism of templated protofilament assembly is sometimes observed during fibril growth. Although the elongation is accomplished mainly by dock-lock monomer addition at the

growing ends, the formation of an ordered protofilament can occur at the lateral surface of a protofibril by collective interconversion of a file of previously deposited monomers. This mechanism is particularly frequent for the model with low β-aggregation propensity, where, due to the frustration of the conformational landscape, the isomerization of a single monomer is strongly disfavored.

In conclusion, the simulation results provide strong evidence of multiple routes of polypeptide self-assembly. Notably, a reduction of the intrinsic β-aggregation propensity induces higher pathway heterogeneity and on-pathway protofibrillar intermediates. Given the experimental evidence of toxicity of prefibrillar aggregates, one is tempted to speculate that therapeutic strategies aimed at reducing fibril-formation propensity (e.g. stabilization of the folded state by small molecules) might paradoxically promote the accumulation of toxic species.

Acknowledgements

We thank F. Marchand and M. Convertino for interesting discussions, and S. Muff for comments on the manuscript. The simulations were performed on the Matterhorn cluster of the University of Zurich, and we gratefully acknowledge the support of C. Bolliger and A. Godknecht. This work was supported by a Swiss National Science Foundation grant and the National Competence Center for Research (NCCR) in Neural Plasticity and Repair.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.09.090

References

1. Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, **426**, 884–890.
2. Lansbury, P. T. & Lashuel, H. A. (2006). A century-old debate on protein aggregation and neurodegeneration enters the clinic. *Nature*, **443**, 774–779.
3. Haass, C. & Selkoe, D. J. (2007). Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid β-peptide. *Nature Rev. Mol. Cell Biol.* **8**, 101–112.
4. Caughey, B. & Lansbury, P. T. (2003). Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. *Annu. Rev. Neurosci.* **26**, 267–298.
5. Nilsberth, C., Westlind-Danielsson, A., Eckman, C. B., Condron, M. M., Axelman, K., Forsell, C. *et al.* (2001). The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Abeta protofibril formation. *Nature Neurosci.* **4**, 887–893.
6. Cheng, I. H., Palop, J. J., Esposito, L. A., Bien-Ly, N., Yan, F. & Mucke, L. (2004). Aggressive amyloidosis in mice expressing human amyloid peptides with the Arctic mutation. *Nature Med.* **10**, 1190–1192.
7. Conway, K. A., Lee, S. J., Rochet, J. C., Ding, T. T., Williamson, R. E. & Lansbury, P. T. (2000). Acceleration of oligomerization, not fibrillization, is a shared property of both alpha-synuclein mutations linked to early-onset Parkinson's disease: implications for pathogenesis and therapy. *Proc. Natl Acad. Sci. USA*, **97**, 571–576.
8. Lomakin, A., Chung, D. S., Benedek, G. B., Kirschner, D. A. & Teplow, D. B. (1996). On the nucleation and growth of amyloid β-protein fibrils: detection of nuclei and quantitation of rate constants. *Proc. Natl Acad. Sci. USA*, **93**, 1125–1129.
9. O'Nuallain, B., Shivaprasad, S., Kheterpal, I. & Wetzel, R. (2005). Thermodynamics of Aβ(1-40) amyloid fibril elongation. *Biochemistry*, **44**, 12709–12718.
10. Kelly, J. W. (1998). The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8**, 101–106.
11. Gosal, W. S., Morten, I. J., Hewitt, E. W., Smith, D. A., Thomson, N. H. & Radford, S. E. (2005). Competing pathways determine fibril morphology in the self-assembly of β2-microglobulin into amyloid. *J. Mol. Biol.* **351**, 850–864.
12. Jahn, T. R., Parker, M. J., Homans, S. W. & Radford, S. E. (2006). Amyloid formation under physiological conditions proceeds *via* a native-like folding intermediate. *Nature Struct. Mol. Biol.* **13**, 195–201.
13. Lomakin, A., Teplow, D. B., Kirschner, D. A. & Benedek, G. (1997). Kinetic theory of fibrillogenesis of amyloid β-protein. *Proc. Natl Acad. Sci. USA*, **94**, 7942–7947.
14. Massi, F. & Straub, J. E. (2001). Energy landscape theory for Alzheimer's amyloid β-peptide fibril elongation. *Proteins: Struct. Funct. Bioinformatics*, **42**, 217–229.
15. Modler, A. J., Gast, K., Lutsch, G. & Damaschun, G. (2003). Assembly of amyloid protofibrils *via* critical oligomers–a novel pathway of amyloid formation. *J. Mol. Biol.* **325**, 135–148.
16. Hall, D., Hirota, N. & Dobson, C. M. (2005). A toy model for predicting the rate of amyloid formation from unfolded protein. *J. Mol. Biol.* **351**, 195–205.
17. Carulla, N., Caddy, G. L., Hall, D. R., Zurdo, J., Gairi, M., Feliz, M. *et al.* (2005). Molecular recycling within amyloid fibrils. *Nature*, **436**, 554–558.
18. Broglia, R. A., Tiana, G., Pasquali, S., Roman, H. E. & Vigezzi, E. (1998). Folding and aggregation of designed proteins. *Proc. Natl Acad. Sci. USA*, **95**, 12930–12933.
19. Gupta, P., Hall, C. K. & Voegler, A. C. (1998). Effect of denaturant and protein concentrations upon protein refolding and aggregation: a simple lattice model. *Protein Sci.* **7**, 2642–2652.
20. Harrison, P. M., Chan, H. S., Prusiner, S. B. & Cohen, F. E. (1999). Thermodynamics of model prions and its implications for the problem of prion protein folding. *J. Mol. Biol.* **286**, 593–606.
21. Dima, R. I. & Thirumalai, D. (2002). Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. *Protein Sci.* **11**, 1036–1049.
22. Urbanc, B., Cruz, L., Yun, S., Buldyrev, S. V., Bitan, G., Teplow, D. B. & Stanley, H. E. (2004). In silico study of amyloid β-protein folding and oligomerization. *Proc. Natl Acad. Sci. USA*, **101**, 17345–17350.
23. Nguyen, H. D. & Hall, C. K. (2004). Molecular

dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl Acad. Sci. USA*, **101**, 16180–16185.
24. Jang, H., Hall, C. K. & Zhou, Y. (2004). Assembly and kinetic folding pathways of a tetrameric beta-sheet complex: molecular dynamics simulations on simplified off-lattice protein models. *Biophys. J.* **86**, 31–49.
25. Khare, S. D., Ding, F., Gwanmesia, K. N. & Dokholyan, N. V. (2005). Molecular origin of polyglutamine aggregation in neurodegenerative diseases. *PLoS Comput. Biol.* **1**, 230–235.
26. Chen, Y. & Dokholyan, N. V. (2005). A single disulfide bond differentiates aggregation pathways of beta2-microglobulin. *J. Mol. Biol.* **354**, 473–482.
27. Malolepsza, E., Boniecki, M., Kolinski, A. & Piela, L. (2005). Theoretical model of prion propagation: a misfolded protein induces misfolding. *Proc. Natl Acad. Sci. USA*, **102**, 7835–7840.
28. Bellesia, G. & Shea, J.-E. (2007). Self-assembly of beta-sheet forming peptides into chiral fibrillar clusters. *J. Chem. Phys.* **126**, 245104.
29. Plakoutsi, G., Bemporad, F., Calamai, M., Taddei, N., Dobson, C. M. & Chiti, F. (2005). Evidence for a mechanism of amyloid formation involving molecular reorganisation within native-like precursor aggregates. *J. Mol. Biol.* **351**, 910–922.
30. Ma, B. & Nussinov, R. (2002). Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers ($A\beta_{16-22}$, $A\beta_{16-35}$, and $A\beta_{10-35}$): sequence effects. *Proc. Natl Acad. Sci. USA*, **99**, 14126–14131.
31. Gsponer, J., Haberthür, U. & Caflisch, A. (2003). The role of side-chain interactions in the early steps of aggregation: molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl Acad. Sci. USA*, **100**, 5154–5159.
32. Klimov, D. & Thirumalai, D. (2003). Dissecting the assembly of $A\beta_{16-22}$ amyloid peptides into antiparallel β sheets. *Structure*, **11**, 295–307.
33. Wei, G., Mousseau, N. & Derreumaux, P. (2004). Sampling the self-assembly pathways of KFFE hexamers. *Biophys. J.* **87**, 3648–3656.
34. Hwang, W., Zhang, S., Kamm, R. D. & Karplus, M. (2004). Kinetic control of dimer structure formation in amyloid fibrillogenesis. *Proc. Natl Acad. Sci. USA*, **101**, 12916–12921.
35. Buchete, N.-V., Tycko, R. & Hummer, G. (2005). Molecular dynamics simulations of Alzheimer's beta-amyloid protofilaments. *J. Mol. Biol.* **353**, 804–821.
36. Lopez de la Paz, M., de Mori, G. M. S., Serrano, L. & Colombo, G. (2005). Sequence dependence of amyloid fibril formation: insights from molecular dynamics simulations. *J. Mol. Biol.* **349**, 583–596.
37. Pellarin, R. & Caflisch, A. (2006). Interpreting the aggregation kinetics of amyloid peptides. *J. Mol. Biol.* **360**, 882–892.
38. Kodali, R. & Wetzel, R. (2007). Polymorphism in the intermediates and products of amyloid assembly. *Curr. Opin. Struct. Biol.* **17**, 48–57.
39. Murphy, R. M. (2007). Kinetics of amyloid formation and membrane interaction with amyloidogenic proteins. *Biochim. Biophys. Acta*, **1768**, 1923–1934.
40. Rao, F. & Caflisch, A. (2004). The protein folding network. *J. Mol. Biol.* **342**, 299–306.
41. Caflisch, A. (2006). Network and graph analyses of folding free energy surfaces. *Curr. Opin. Struct. Biol.* **16**, 71–78.
42. Esler, W. P., Stimson, E. R., Jennings, J. M., Vinters, H. V., Ghilardi, J. R., Lee, J. P. *et al.* (2000). Alzheimer's disease amyloid propagation by a template-dependent dock-lock mechanism. *Biochemistry*, **39**, 6288–6295.
43. Gobbi, M., Colombo, L., Morbin, M., Mazzoleni, G., Accardo, E., Vanoni, M. *et al.* (2006). Gerstmann-Sträussler-Scheinker disease amyloid protein polymerizes according to the "dock-and-lock" model. *J. Biol. Chem.* **281**, 843–849.
44. Nguyen, P. H., Li, M. S., Stock, G., Straub, J. E. & Thirumalai, D. (2007). Monomer adds to preformed structured oligomers of abeta-peptides by a two-stage dock-lock mechanism. *Proc. Natl Acad. Sci. USA*, **104**, 111–116.
45. Jansen, R., Dzwolak, W. & Winter, R. (2005). Amyloidogenic self-assembly of insulin aggregates probed by high resolution atomic force microscopy. *Biophys. J.* **88**, 1344–1353.
46. Walsh, D., Hartley, D. M., Kusumoto, Y., Fezoui, Y., Condron, M., Lomakin, A. *et al.* (1999). Amyloid β-protein fibrillogenesis. structure and biological activity of protofibrillar intermediates. *J. Biol. Chem.* **274**, 25945–25952.
47. Fersht, A. (1999). *Structure and Mechanism in Protein Science.* W.H. Freeman and Company, New York, NY.
48. Tartaglia, G. G., Cavalli, A., Pellarin, R. & Caflisch, A. (2004). The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.* **13**, 1939–1941.

Supplementary Materials

Pathways and intermediates of amyloid fibril formation

Supplementary Information

Riccardo Pellarin, Enrico Guarnera, and Amedeo Caflisch*

Department of Biochemistry, University of Zürich,
Winterthurerstrasse 190, CH-8057 Zürich, Switzerland
email: caflisch@bioc.unizh.ch, Phone: +41 44 635 55 21, FAX: +41 44 635 68 62,
**Corresponding author*

Contents

1 Kinetics of the dock-lock mechanism 2

2 Kinetic traces of β-unstable and β-stable models 5

1 Kinetics of the dock-lock mechanism

It has been reported that the deposition of monomers onto the fibril follows a dock-lock mechanism (1, 2), which is termed templated assembly elsewhere (3, 4). According to this mechanism, monomeric peptide diffuses in solution to the fibril end and then undergoes a conformational reorganization to the locked state.

To evaluate the efficiency of the dock-lock mechanism five states s_i were defined for the monomer: the isolated π-state 1_π, the isolated β-state 1_β, the fibrillar π-state F_π, the fibrillar β-state F_β and a state s_0 containing all the remaining aggregates. These states determine whether a monomer is either isolated or assembled to a fibril, and if it is in the β or π-state. The fibril is defined as an aggregate with more than 50 monomers. Here we are interested in the association kinetics of monomers onto the fibril at the steady-state, where the elongation phase has ended and the mature fibril is in equilibrium with monomers. Below we evaluate the rates of interconversion using the Markov chains formalism.

The transitions occurred in the 100 trajectories of models $dE = -1.5$ and $dE = -2.5$ kcal/mol have been collected and analyzed, and the transition matrix between the aforementioned states was constructed. The transition matrix $\mathbf{T}$ at the time interval $\tau = 0.5$ ns is estimated as

$$T(s_j|s_i;\tau) = \frac{P(s_i \cap s_j;\tau)}{w(s_i)} \tag{1}$$

where $P(s_i \cap s_j;\tau)$ is the probability flow, calculated on the transitions occurred between the states s_i and s_j within the time τ, and $w(s_i)$ is the equilibrium probability of the state s_i. The equilibrium probabilities are given by $w(s_i) = n(s_i)/\nu$ where $n(s_i)$ is the number of occurrences of s_i, and ν is the length of the simulation in frames. The probability flow is $P(s_i \cap s_j;\tau) = n(s_i \cap s_j;\tau)/(\nu - \nu_s)$ where $n(s_i \cap s_j;\tau)$ is the number of transitions $s_i \to s_j$ and ν_s the number of independent simulations. Thus the entries of the 5×5 matrix $\mathbf{T}$ are conditional probabilities for a single microscopic transition. Using the Markov chain formalism, the equilibrium inter-conversion rates were estimated from the mean first passage time matrix **MFPT**. $\mathbf{MFPT}(s_i \to s_j)$ is the average time, starting at s_i, needed to reach the state s_j for the first time, and can be derived from the transition matrix $\mathbf{T}$ exploiting the ergodicity of the relative Markov chain (5, 6). The procedure is based on elementary linear algebra and requires to define a fundamental matrix for ergodic Markov chains, $\mathbf{Z}$:

$$\mathbf{Z} = (\mathbf{I} - \mathbf{T} + \mathbf{W})^{-1} \tag{2}$$

where $\mathbf{I}$ is the identity matrix, and the matrix of equilibrium populations $\mathbf{W}$ is

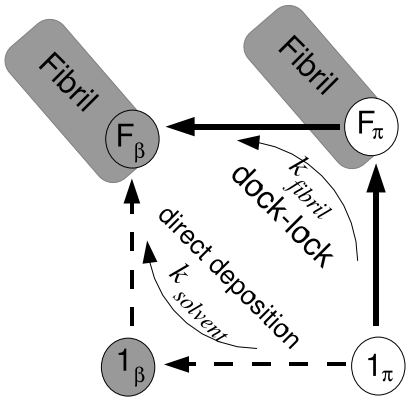

Fig. 1. Routes of association. In the "deposition and reorganization", the isolated monomer in the π-state first deposits at the fibril surface, and afterwards converts to the β-state. In the "direct deposition", the isolated monomer in the π-state first converts to the β-state, and then associate to the fibril.

defined as $W_{ij} = w(s_j)$ for each i. The matrix **MFPT** is eventually given by the expression

$$\mathbf{MFPT} = \tau(\mathbf{I} - \mathbf{Z} + \mathbf{E}\mathbf{Z}_{dg})\mathbf{D} \tag{3}$$

D is a diagonal matrix with element $1/w(s_j)$ at the jth diagonal entry, **E** is a matrix whose elements are all 1's, and $\mathbf{Z}_{dg}$ contains the diagonal elements of **Z**. The factor τ converts the time units from simulation frames to nanoseconds. The inverse of the **MFPT** matrix elements gives the macroscopic rate of the transition between a state s_i to a state s_j:

$$k(s_i \to s_j) = \frac{1}{\mathbf{MFPT}(s_i \to s_j)} \tag{4}$$

The rate of a two step process is derived as follows:

$$k(s_i \to s_j \to s_l) = \frac{k(s_i \to s_j)k(s_j \to s_l)}{k(s_i \to s_j) + k(s_j \to s_l)} \tag{5}$$

As explained in the main text, the two possible routes of association are the "deposition and reorganization", which is the dock-lock mechanism, and the "direct deposition" (see Fig. 1). The former corresponds to the transition $1_\pi \to F_\pi \to F_\beta$

and has a rate k_{fibril}, and the latter to $1_\pi \rightarrow 1_\beta \rightarrow F_\beta$ with a rate $k_{solvent}$. Using Eq. 4 and Eq. 5 we obtained for $dE = -2.5$ kcal/mol: $k_{fibril} = 1.2 \cdot 10^{-2}$ ns^{-1} and $k_{solvent} = 3.2 \cdot 10^{-3}$ ns^{-1}, while for $dE = -1.5$ kcal/mol: $k_{fibril} = 2.5 \cdot 10^{-2}$ ns^{-1} and $k_{solvent} = 3.8 \cdot 10^{-3}$ ns^{-1}.

References

1. Esler, W. P., Stimson, E. R., Jennings, J. M., Vinters, H. V., Ghilardi, J. R., Lee, J. P., Mantyh, P. W. & Maggio, J. E. (2000). Alzheimer's disease amyloid propagation by a template-dependent dock-lock mechanism. *Biochemistry,* **39** (21), 6288–6295.
2. Gobbi, M., Colombo, L., Morbin, M., Mazzoleni, G., Accardo, E., Vanoni, M., Favero, E. D., Cantú, L., Kirschner, D. A., Manzoni, C., Beeg, M., Ceci, P., Ubezio, P., Forloni, G., Tagliavini, F. & Salmona, M. (2006). Gerstmann-Sträussler-Scheinker disease amyloid protein polymerizes according to the "dock-and-lock" model. *J Biol Chem,* **281** (2), 843–849.
3. Griffith, J. S. (1967). Self-replication and scrapie. *Nature,* **215** (5105), 1043–1044.
4. Serio, T. R., Cashikar, A. G., Kowal, A. S., Sawicki, G. J., Moslehi, J. J., Serpell, L., Arnsdorf, M. F. & Lindquist, S. L. (2000). Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science,* **289** (5483), 1317–1321.
5. Snell, J. (1959). Finite Markov Chains and their Applications. *The American Mathematical Monthly,* **66** (2), 99–104.
6. Kemeny, J. & Snell, J. (1976). *Finite Markov Chains*. Springer.

2 Kinetic traces of β-unstable and β-stable models

Fig. 2. Twenty-five time series of the 100 runs of the β-unstable model ($dE = -2.5$ kcal/mol): size of the largest aggregate N_{la} (red curve), the number of monomers in the β-state N_{la}^{β} (green curve) and the number of protofibrils N_{la}^{pf} (black curve). For N_{la} and N_{la}^{β} the y-values have been scaled by a factor 0.1. The x-axis has the units of μs.

Fig. 3. Twenty-five time series of the 100 runs of the β-stable model ($dE = -1.5$ kcal/mol). For description of the curves see Fig. 2. Since in the lag phase two fibrils can nucleate within the simulation box for this model, the number of protofilaments per fibril can temporarily be larger than 4 when the two fibrils merge together in the elongation phase. This event is accompanied by an abrupt increase of both N_{la} and N_{la}^{β}.

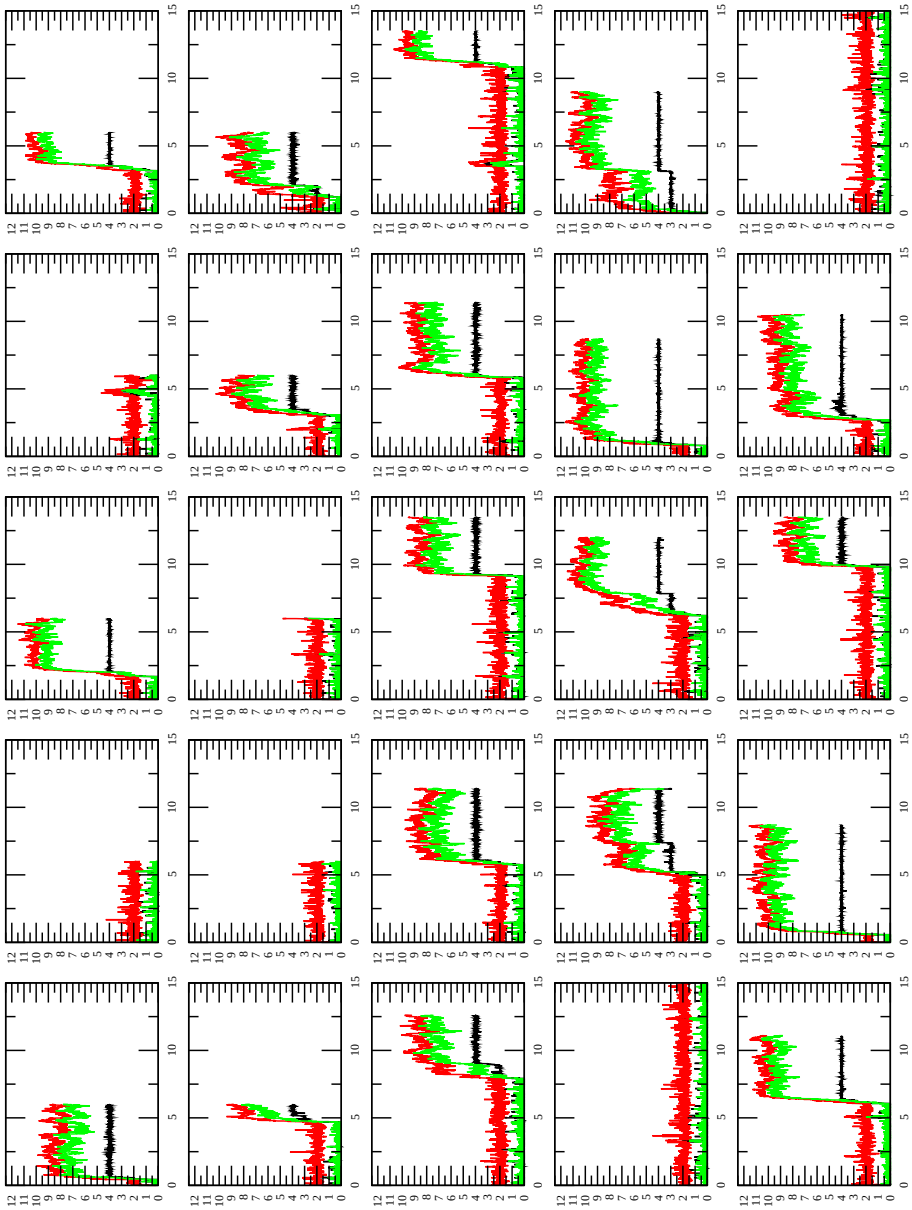

Fig. 1.

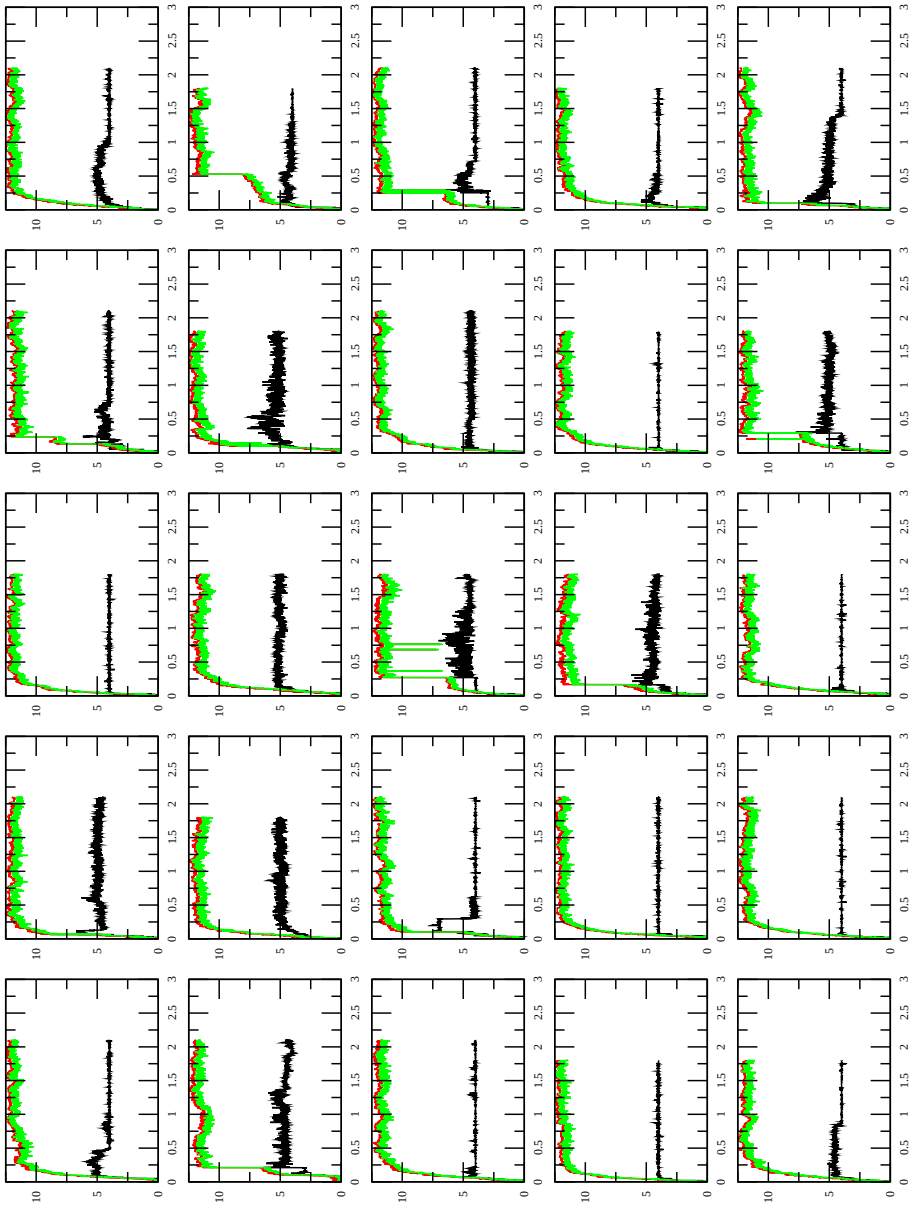

Fig. 2.

7 Conclusions and outlook

The protein folding problem is one of the most intriguing challenges of the modern experimental and theoretical biophysics. Its solution in globular proteins not only means to uncover the link between the sequence of amino acids and the three dimensional structure at which proteins exert their function. It also means to explain what are the hidden mechanisms through which a polypeptide can easily reach its lowest free energy configuration. The problem is inherently complex due to the large number of degrees of freedom into play. Since only phenomenological theories exist about complex systems a crucial aspect of the investigations is that of the "right description", the right choice of the key variables which allows the researcher to drow conclusions out of the observations. That is especially true in computer simulations where the extraction of the relevant information is often arbitrary. Computer simulations of proteins, and more in general of bio-molecules, have conquered a central role on the field, very close to that played by the real world experiments. As much as experiments they produce data, Gbytes of data that need to analyzed, interpreted, assessed [Larson et al., 2002]. In this thesis one of the topics investigated is that of the description of the configurational space of proteins with the aim to provide a framework for the analysis of the complex folding dynamics. The estimated thermodynamic parameters from a simulation strongly depend on the definition of coarse-grained states or mesoscopic descriptions. We have learned that if a mesocopic description generates a discrete and finite partitioning of the configurational states, the total entropy of the system is split into an informational term plus a term representing the internal vibrational modes of the mestostates: $S = H + S^{\mathrm{b}} = \sum_i^N P_i(H_i + S_i^{\mathrm{b}})$ where $H_i = -k_B \ln P_i$ is the information per mesostate of probability P_i and $S_i^{\mathrm{b}} = \int \rho_i(\Gamma) \ln \rho_i(\Gamma) d\Gamma$ the internal entropy per mesostate with $\rho_i(\Gamma)$ the Gibbs distribution within a partition i. If the mesoscopic description is multidimensional, such as the SRA[4] (strings of rotational angles) the informational part can be interpreted as configurational entropy. Thus the thermodynamic stability of the states can be evaluated on two levels, that one proper of the vibrational modes and that one as a consequence of the partitioning. On the basis of this observation one can evaluate on pure informational basis the "quality" of a mesoscopic description. The good mesoscopic descriptions are those maximizing the amount of information extracted from the ensemble of microstates. Since computer simulations are finite in time, finite size effects affect the estimation of the thermodynamic parameters. Each mesostate is defined by its observed probability P_i and by its free energy content $G_i = \overline{E_i} - TS_i^{\mathrm{b}}$. We found that the distribution of the G_i values, in the context of a particular mesoscopic description, is always double peaked, one corresponding to positive G_i values and the other to negative values (see figure 2.4). The positive peak is always gaussian and corresponds to ensemble of mesostates with a poor statistics (unstable states) but large in number while the other narrow peak reflects mesostates that are well sampled (stable states) but few in number, about the 10 %. The analysis conducted on the GSGS peptide showed that there are only few few stable configurational mesostates among which the peptide dynamics takes place. This means that the configurational space of foldable proteins is not only divided in a folded and an unfolded state.

Within the unfolded state the multiple presence of low free energy states suggests that the search for the folded state is not just trial and error, but it is rather driven by a limited set of conformers. It has been argued that the organization of the configurational space of proteins as a direct consequence of the organization of their free energy landscape [Frauenfelder and Leeson, 1998, Frauenfelder, 2002, Krivov and Karplus, 2002, Caflisch, 2006] is the product of an evolutionary development. The mesoscopic description of the configurational space by means of multidimensional strings of symbols is particularly appropriated for proteins. Proteins are essentially *digital* objects, the sequence of amino acids direct translation of the genetic code is a digital object. The information fully transfered from the genetic code to proteins through translation, is a intriguing example of digital transmission of information. It is reasonable to think that a similar digital communication channel exists between the sequence of amino acids and the companion configurational space. In this sense we do believe that a symbolic description of the configurational space is not only useful but also necessary to fully comprehend the relations sequence/structure in proteins. Thus if a symbolic approach is employed in the treatment of a physical system, then combining statistical mechanics and information theory is an opportune step [Crofts, 2007].

The investigation of protein folding kinetics through molecular dynamics simulations introduces the problem of which reaction coordinate adopting to fully characterize folding. Here we have proposed a strategy analysis based on Markov chains which is on the line with previous works on the subject [Swope et al., 2004a, Chodera et al., 2007, Li et al., 2008]. With our method the ensemble of microstates generated by the molecular dynamics simulations are first discretized in mesostates (either through clustering or digital symbolization) and redefined in terms of their causality kinetic properties. We introduced a method called causal grouping according to which a minimal Markovian master equation can built up to fully describe folding kinetics. Causal grouped states are defined through a redistribution of mesostates that are statistically not meaninful to those statistically stable by using causal dynamical connectivity. The strategy does not require any assumption on the diffusive properties of the configurational space. This method allows to monitor folding kinetics at mesoscopic level, through the analysis of the transition matrix, and at equilibrium through the evolution of the Markov process. The analysis at mesoscopic level consists in the network representation on the transition matrix that in case of a Markovian description represents the ensemble of possible mesoscopic pathways a molecule can cover (see figure 2.30). The equilibrium level is given by the matrix of the mean first passage times MFPT matrix between all couples of causal mesostates. The MFPT matrix contains all the equilibrium mean free energy barriers between all the mesostates into play, in particular it gives insight on the macroscopic mechanisms of the folding reaction (see figure 2.32). From the MFPT matrix a one-dimensional free energy profiles can be as well extracted (see figure 3.14) which provide a quantitative kinetic overview of how the overall barriers separate unfolded macrostates from folded. The MFPT matrix applied to the GSGS folding dynamics revealed that, unlike to what one could guess, the dynamics in the unfolded state is barely diffusive. The basins characterizing the unfolded state do exchange between them on a time scale greater than the folding time, which means that the folded state represents the "hub" that allows communication among different basins. In other words when the system stands in an unfolded basin it prefers to jump into the folded state in one go instead diffusing through other unfolded basins. This result suggests a pre-organized picture of the unfolded state, made of independent basins that might represent the gateways to the folding routes. A "star" shaped configurational space with the folded state occupying the centre place reconciles the experimental findings on folding kinetics, and the results

from simulations. The unfolded basins organized like in a ring surrounding the folded state, and barely communicating between them, can give arise to an overall folding free energy barrier which explains a two state behavior of most foldable proteins. These results also suggest an organization of the configurational space consistent with a hierarchical tree [Rose et al., 2006, Hockenmaier et al., 2007, Ozkan et al., 2007] in which the local order of the unfolded state define the gateways to productive folding routes.

The method of the Markovian treatment of complex dynamics has been applied also to the description of the fibril formation pathways in the context of coarse-grained simulations of polypeptide aggregation. The coarse-grained polypeptides are characterized by a free energy profile having a distinct amyloid-competent (i.e. beta-prone) state and an amyloid-protected state [Pellarin and Caflisch, 2006, Pellarin et al., 2007]. A decrease in the β-aggregation propensity resulted in a larger heterogeneity of elongation pathways, despite the essentially identical structure of the final fibril. Thus according to to these simulations if the β-aggregation propensity is high the fibrils are formed through a simple deposition mechanism: the monomers first change state from aggregation-protected to β-prone and later rapidly polymerize until the mature fibril is formed; conversely if the β-aggregation propensity is low the system shows a lag phase during which monomers coordinate themselves into micelles and then, after the formation of critical nucleus, a proliferation of diverse pathways drive the system to the mature fibril. Thanks to our method we could quantitatively estimate what are the preferential pathways that leads to the fiber elongation. The use of a master equation for these systems is highly facilitated by the natural definition of state in terms of monomers and aggregation numbers. In the future we plan to fully exploit the potentialities of the master equation to uncover all the dominant pathways in the aggregation process: from the formation of critical nucleus to the coordination of stable diffusive oligomeric species and their overall interplay in the aggregation reaction.

The method of the Markovian description of time series together with the construction of causal states is quite powerful and in principle can be applied to any kind of time series of a stochastic process. Appealing applications of this method are the single molecule time series from FRET and ET experiments [Talaga et al., 2000, Schuler et al., 2002, Lipman et al., 2003, Haran, 2003, Yang et al., 2003, Neuweiler and Sauer, 2004]. Attempts on this line are carried on by using the computational mechanics framework [Shalizi and Crutchfield, 2002, Li et al., 2008] although the definition of dynamic state in single molecule data is not free from ambiguities.

Another important topic investigated in this thesis was the study of simplified protein by means of molecular dynamics simulations. Simplified protein sequences were constructed by using an amino acid alphabet of solely three letters. The aim of the study was twofold. First we made the hypothesis that for certain protein topologies low complexity amino acid alphabets were able to encode, although in shallow manner, the overall structural properties of the folded states. Implicitly such an hypothesis implies that the evolutionary pattern that generated the modern protein sequences was driven towards the specialization of protein functions rather than protein structures. Secondly, if the first hypothesis is true how can we simplify protein sequences such that their folding mechanisms can be observed in a molecular dynamics simulation. No matter how accurate can be the modern force fields, nowadays only short peptides are accessible for all atom folding studies with computer simulations. Even if force fields are accurate enough to predict a folded state as lowest free energy state the computational time to prove that might not be possible. Our question was: can we simplify a small size (say about 60 residues) protein sequence so that the folding rate is increased enough to observe spontaneous folding in silico?

The amino acids for the simplified alphabet were chosen according to their secondary structure propensities. We construct five proteins of two kind: four full β-sheet proteins of respectively 20, 28, 36, 44 residues and a α/β protein of 56 residues. The α/β protein sequence is a simplified version of the B1 domain of protein G. In equilibrium molecular dynamics simulations we observed reversible folding for all proteins studied. Most importantly the observed folded states corresponded to those that the sequence design was aimed for, notably the folded state of the α/β simplified protein resulted structurally very close to that of protein G. The folded states of these proteins are liquid like, characterized by a persistent secondary structure, the lack of specific tertiary contacts and hydrophobic core. The hydrophobic effect is not the main driving force for the folding of these proteins. Folding initiates from the highly disordered turns which facilitate the coordination of β-hairpins by lowering the entropic barrier for their formation. The folded states are marginally stable of about 1 kcal/mol but highly accessible from the kinetic point of view. The configurational space of the α/β protein is characterized by an extraordinary low dynamic frustration and is populated by a large variety of different configurational states. Notably the presence of fiber-like states play the role of kinetic traps (see figure 3.13). This fact suggests that the role of evolution in increasing protein function efficiency included also the duty to select sequences with a low propensity toward potentially dangerous configurational states. Thus if on one hand simplified sequence can encode ordered folded structures and generate low frustrated energy landscapes, but on the other hand they give kinetic access to misfolded states. The application of the causal grouped description for the investigation of the protein dynamics showed a folding mechanism not very dissimilar from that observed for the GSGS peptide. Also in this case, for instance the α/β simplified protein, the folded state plays the role of a central connector between all the free energy basins. In the unfolded state the basins are independent and connected to the folded state trough a single jump. As experiments have also demonstrated [Davidson and Sauer, 1994, Davidson et al., 1995, Riddle et al., 1997] simplified sequence can be a clue to study the folding problem. We believe that low complexity amino acid alphabets can encode complicated protein topologies when the underling sequence punctuation is fully understood. Our simulations showed for example the double role played by the residues responsible for the turns. In the unfolded state they favor the search of secondary contacts by lowering the entropic barrier while in the folded state they act as entropic stabilizers. Experiments to verify whether our predictions are correct of the α/β protein are in the course. To overcome possible solubility issues due to the anphiphilic character of the simplified sequence the insertion of few polar residues (such lysines) might reduce the severity of problem. A very positive response from these experiments would be the observation a molten globular state by means for instance of ANS binding experiment. It would be nice to apply the same simplification scheme to study other proteins such as the fully computationally designed α/β Top7 protein [Kuhlman et al., 2003], or more in general protein topologies that are mainly stabilized by secondary contacts.

As a final remark we would like to show the evolution of the literature production along the last 40 years. On the ISI web of science we searched for all the publications having “protein folding” as a research topic. The result of this search is interesting and is shown in figure 7.1. In figure we plotted the number of publications issued per year (black curve), the mean number of citations (blue curve) that the papers issued in a certain year received (and are still receiving), and the number of citations that the most cited work (red curve) received (and is still receiving). The peaks in the blue and red curve show how intensively productive was a year, namely how important was the research in that year. A high cited

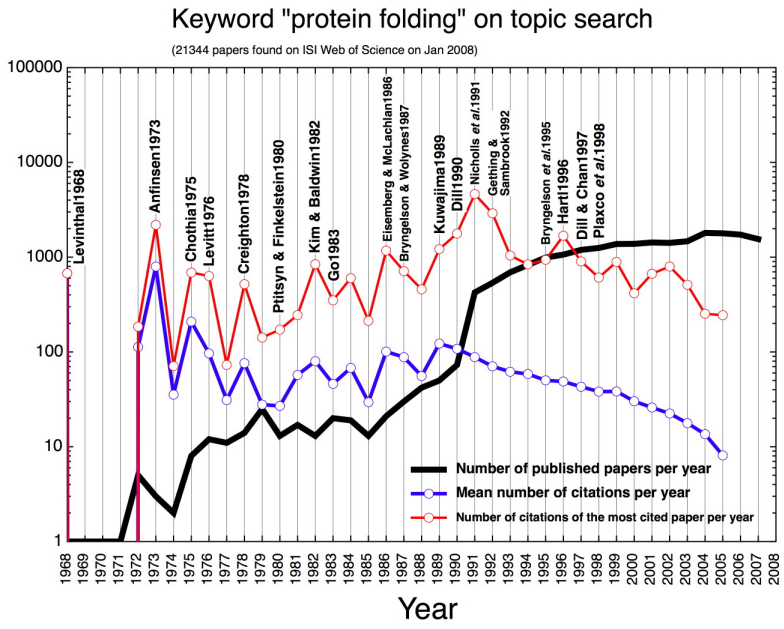

Figure 7.1: The scientific production on the "protein folding" topic of the last 40 years. The black curve represents the total number of published articles per year: after a constant increase from 1973 to the end of the 80s a bursting increase characterize the beginning of the 90s, which is followed by a plateau. The blue curve is the mean number of citations of the papers published in a certain year which is calculated from the total number of citations per year divided by the total number of published papers per year. The red curve gives the number of citations corresponding to the most cited paper per year: peaks corresponds to outstanding papers that represent a breakthroughs in the field (some representative citations are included). The blue and red curves in the late tails are clearly affected by an incomplete statistics. Do the trends suggest an imminent decline of protein folding as an autonomous research topic or the coming of new breakthroughs?

work represents a breakthrough in the field which generates also a scientific inheritance. High values of the mean number of citations tells how diffuse is the future importance of the scientific production in a specific year. The black curve, the total production of papers, after a bursting phase at the end of the 80s and beginning of the 90s (these were the years of the molten globule discovery, of the energy landscape perspective and of the studies on *in vivo* protein folding) from about the beginning of the new century a plateau seem to be reached, and from the 2007 data it seems that a decreasing tendency for the first time has started. What might that mean? In economy we know that when an indicator reach a steady state it will soon either fall or growing. We strongly hope for the second option, the time for new challenges and new perspectives in protein folding has arrived. ♣

Bibliography

[Abundo et al., 2002] Abundo, M., Accardi, L., Rosato, N., and Stella, L. (2002). Analysing protein energy data by a stochastic model for cooperative interactions: comparison and characterization of cooperativity. *J Math Biol*, 44(4):341–359.

[Alberts et al., 1998] Alberts, B. et al. (1998). *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*. Garland New York:.

[Ananthanarayanan et al., 1984] Ananthanarayanan, V. S., Brahmachari, S. K., and Pattabiraman, N. (1984). Proline-containing beta-turns in peptides and proteins: analysis of structural data on globular proteins. *Arch Biochem Biophys*, 232(2):482–95.

[Andersen et al., 2002] Andersen, C. A. F., Palmer, A. G., Brunak, S., and Rost, B. (2002). Continuum secondary structure captures protein flexibility. *Structure*, 10(2):175–184.

[Anfinsen, 1973] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(96):223–230.

[Arai and Kuwajima, 1996] Arai, M. and Kuwajima, K. (1996). Rapid formation of a molten globule intermediate in refolding of alpha-lactalbumin. *Fold Des*, 1(4):275–87.

[Auber, 2003] Auber, D. (2003). Tulip : A huge graph visualisation framework. In Mutzel, P. and Jünger, M., editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag.

[Badii and Politi, 1999] Badii, R. and Politi, A. (1999). *Complexity: Hierarchical Structures and Scaling in Physics*. Cambridge University Press.

[Baldwin, 2007] Baldwin, R. L. (2007). Energetics of protein folding. *J Mol Biol*, 371(2):283–301.

[Baldwin and Rose, 1999a] Baldwin, R. L. and Rose, G. D. (1999a). Is protein folding hierarchic? i. local structure and peptide folding. *Trends in Biochemical Sciences*, 24(1):26–33.

[Baldwin and Rose, 1999b] Baldwin, R. L. and Rose, G. D. (1999b). Is protein folding hierarchic? ii. folding intermediates and transition states. *Trends in Biochemical Sciences*, 24(2):77–83.

[Berkenpas et al., 1995] Berkenpas, M. B., Lawrence, D. A., and Ginsburg, D. (1995). Molecular evolution of plasminogen activator inhibitor-1 functional stability. *EMBO J*, 14(13):2969–2977.

[Berman et al., 2003] Berman, H., Westbrook, J., Zardecki, C., and Bourne, P. (2003). The Protein Data Bank. *Protein Structure: Determination, Analysis, and Applications for Drug Discovery*.

[Best and Hummer, 2005] Best, R. B. and Hummer, G. (2005). Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA*, 102(19):6732–6737.

[Betz and De Grado, 1996] Betz, S. and De Grado, W. (1996). Controlling topology and native-like behavior of de novo-designed peptides: Design and characterization of antiparallel four-stranded coiled coils. *Biochemistry*, 35:6955–6962.

[Bonetto and Gallavotti, 1997] Bonetto, F. and Gallavotti, G. (1997). Reversibility, coarse graning and the chaoticity principle. *Commun. Math. Phys.*, 189:263–275.

[Brooks et al., 1983] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM - a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217.

[Bruß and Frick, 1995] Bruß, I. and Frick, A. (1995). Fast interactive 3-d graph visualization. *Proceedings of Graph Drawing'95*, pages 99–110.

[Bryngelson et al., 1995] Bryngelson, J., Onuchic, J., Socci, N., and Wolynes, P. (1995). Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Structure, Function, and Genetics*, 21:167–195.

[Bryngelson and Wolynes, 1987] Bryngelson, J. and Wolynes, P. (1987). Spin Glasses and the Statistical Mechanics of Protein Folding. *Proc. Natl. Acad. Sci. USA*, 84(21):7524–7528.

[Buchler and Goldstein, 1999a] Buchler, N. and Goldstein, R. (1999a). Universal correlation between energy gap and foldability for the random energy model and lattice proteins. *J. Chem. Phys.*, 111:6599–6609.

[Buchler and Goldstein, 1999b] Buchler, N. E. and Goldstein, R. A. (1999b). Effect of alphabet size and foldability requirements on protein structure designability. *Proteins: Structure, Function, and Bioinformatics*, 34(1):113–124.

[Cabrita and Bottomley, 2004] Cabrita, L. and Bottomley, S. (2004). How do proteins avoid becoming too stable? Biophysical studies into metastable proteins. *European Biophysics Journal*, 33(2):83–88.

[Caflisch, 2003] Caflisch, A. (2003). Folding for binding or binding for folding? *Trends In Biotechnology*, 21:423–425.

[Caflisch, 2006] Caflisch, A. (2006). Network and graph analyses of folding free energy surfaces. *Curr. Opin. Struct. Biol.*, 16(1):71–78.

[Caflisch and Paci, 2004] Caflisch, A. and Paci, E. (2004). *Protein folding handbook*, chapter Molecular dynamics simulations to study protein folding and unfolding, pages 1143–1169. Wiley-VCH, Weinheim.

[Cavalli et al., 2002] Cavalli, A., Ferrara, P., and Caflisch, A. (2002). Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins: Structure, Function, and Bioinformatics*, 47(3):305–314.

[Cavalli et al., 2003] Cavalli, A., Haberthür, U., Paci, E., and Caflisch, A. (2003). Fast protein folding on downhill energy landscape. *Prot. Sci.*, 12(8):1801–1803.

[Cavalli et al., 2005] Cavalli, A., Vendruscolo, M., and Paci, E. (2005). Comparison of sequence-based and structure-based energy functions for the reversible folding of a peptide. *Biophys. J.*, 88(5):3158–3166.

[Cecchini et al., 2006] Cecchini, M., Curcio, R., Pappalardo, M., Melki, R., and Caflisch, A. (2006). A molecular dynamics approach to the structural characterization of amyloid aggregation. *J Mol Biol*, 357(4):1306–1321.

[Cecchini et al., 2004] Cecchini, M., Rao, F., Seeber, M., and Caflisch, A. (2004). Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J Chem Phys*, 121(21):10748–10756.

[Chodera et al., 2007] Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A., and Swope, W. C. (2007). Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *J Chem Phys*, 126(15):155101.

[Chothia and Finkelstein, 1990] Chothia, C. and Finkelstein, A. (1990). The Classification and Origins of Protein Folding Patterns. *Annual Review of Biochemistry*, 59(1):1007–1035.

[Cieplak et al., 1998] Cieplak, M., Henkel, M., Karbowski, J., and Banavar, J. R. (1998). Master equation approach to protein folding and kinetic traps. *Phys. Rev. Lett.*, 80:3654–3657.

[Cordes et al., 1996] Cordes, M. H., Davidson, A. R., and Sauer, R. T. (1996). Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.*, 6(1):3–10.

[Creighton, 1997] Creighton, T. (1997). How important is the molten globule for correct protein folding? *Trends in Biochemical Sciences*, 22(1):6–10.

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–3.

[Crofts, 2007] Crofts, A. R. (2007). Life, information, entropy, and time. *Complexity*, 13:14–50.

[Crutchfield and Young, 1989] Crutchfield, J. P. and Young, K. (1989). Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108.

[Dames et al., 2006] Dames, S. A., Aregger, R., Vajpai, N., Bernadó, P., Blackledge, M., and Grzesiek, S. (2006). Residual dipolar couplings in short peptides reveal systematic conformational preferences of individual amino acids. *J. Am. Chem. Soc.*, 128(41):13508–14.

[Daura et al., 1999a] Daura, X., Antes, I., van Gunsteren, W. F., Thiel, W., and Mark, A. E. (1999a). The effect of motional averaging on the calculation of NMR-derived structural properties. *Proteins: Structure, Function, and Bioinformatics*, 36(4):542–555.

[Daura et al., 1999b] Daura, X., van Gunsteren, W. F., and Mark, A. E. (1999b). Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. *Proteins: Structure, Function, and Bioinformatics*, 34(3):269–280.

[Davidson et al., 1995] Davidson, A. R., Lumb, K. J., and Sauer, R. T. (1995). Cooperatively folded proteins in random sequence libraries. *Nature Struct. Biol.*, 2(10):856–864.

[Davidson and Sauer, 1994] Davidson, A. R. and Sauer, R. T. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA*, 91(6):2146–2150.

[Dayalan et al., 2006] Dayalan, S., Gooneratne, N. D., Bevinakoppa, S., and Schroder, H. (2006). Dihedral angle and secondary structure database of short amino acid fragments. *Bioinformation*, 1(3):78–80.

[De Alba et al., 1999] De Alba, E., Santoro, J., Rico, M., and Jiménez, M. A. (1999). De novo design of a monomeric three-stranded antiparallel beta-sheet. *Prot. Sci.*, 8(4):854–865.

[de Groot et al., 2001] de Groot, B. L., Daura, X., Mark, A. E., and Grubmuller, H. (2001). Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. *J Mol Biol*, 309(1):299–313.

[Di Giulio, 1996] Di Giulio, M. (1996). The β-sheets of proteins, the biosynthetic relationships between amino acids, and the origin of the genetic code. *Origins Life Evol B*, 26(6):589–609.

[Di Giulio, 1997] Di Giulio, M. (1997). On the origin of the genetic code. *J Theor Biol*, 187(4):573–81.

[Dill, 1985] Dill, K. (1985). Theory for the folding and stability of globular-proteins. *Biochemistry-Us*, 24:1501–1509.

[Dill, 1990a] Dill, K. (1990a). Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155.

[Dill, 1990b] Dill, K. (1990b). The meaning of hydrophobicity. *Science*, 250:297–297.

[Dill and Chan, 1997] Dill, K. A. and Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Struct. Biol.*, pages 10–19.

[Dill and Shortle, 1991] Dill, K. A. and Shortle, D. (1991). Denatured states of proteins. *Ann. Rev. Biochem.*, 60:795–825.

[Dinner et al., 1999] Dinner, A., Lazaridis, T., and Karplus, M. (1999). Understanding beta-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 96:9068–9073.

[Dinner et al., 2000] Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M., and Karplus, M. (2000). Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biotechn.*, 25:331–339.

[Dobson, 2003] Dobson, C. (2003). Protein folding and misfolding. *Nature*, 426(6968):884–890.

[Dobson, 1999] Dobson, C. M. (1999). Protein misfolding, evolution and disease. *Trends Biotechn.*, 24:329–332.

[Domany, 1999] Domany, E. (1999). Superparamagnetic clustering of data - The definitive solution of an ill-posed problem. *Physica A*, 263:158–169.

[Du et al., 1998] Du, R., Pande, V., Grosberg, A., Tanaka, T., and Shakhnovich, E. (1998). On the transition coordinate for protein folding. *J. Chem. Phys.*, 108(1):334.

[Dunker et al., 2001] Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001). Intrinsically disordered protein. *J Mol Graph Model*, 19(1):26–59.

[Eaton et al., 2000] Eaton, W. A., Munoz, V., Hagen, S. J., Jas, G. S., Lapidus, L. J., Henry, E. R., and Hofrichter, J. (2000). Fast kinetics and mechanisms in protein folding. *Ann. Rev. Biophys. Biomol. Struct.*, 29:327–359.

[Ebeling, 1993] Ebeling, W. (1993). Entropy and information in processes of self-organization: uncertainty and predictability. *Physica A*, 194:563–575.

[Ebeling and Klimontovič, 1984] Ebeling, W. and Klimontovič, J. (1984). *Selforganization and Turbulence in Liquids*. Teubner.

[Fan and Wang, 2003] Fan, K. and Wang, W. (2003). What is the minimum number of letters required to fold a protein? *J. Mol. Biol.*, 328(4):921–926.

[Ferrara et al., 2000] Ferrara, P., Apostolakis, J., and Caflisch, A. (2000). Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J. Phys. Chem. B*, 104(20):5000–5010.

[Ferrara et al., 2002] Ferrara, P., Apostolakis, J., and Caflisch, A. (2002). Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 46:24–33.

[Ferrara and Caflisch, 2000] Ferrara, P. and Caflisch, A. (2000). Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl. Acad. Sci. USA*, 97:10780–10785.

[Fersht, 1995] Fersht, A. (1995). Optimization of rates of protein-folding - the nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci. USA*, 92:10869–10873.

[Fersht, 1997] Fersht, A. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol*, 7(3):9.

[Fersht, 1999] Fersht, A. (1999). *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. WH Freeman New York.

[Fersht and Daggett, 2002] Fersht, A. R. and Daggett, V. (2002). Protein folding and unfolding at atomic resolution. *Cell*, 108(4):573–582.

[Fierz and Kiefhaber, 2007] Fierz, B. and Kiefhaber, T. (2007). End-to-end vs interior loop formation kinetics in unfolded polypeptide chains. *J Am Chem Soc*, 129(3):672–9.

[Fierz et al., 2007] Fierz, B., Satzger, H., Root, C., Gilch, P., Zinth, W., and Kiefhaber, T. (2007). Loop formation in unfolded polypeptide chains on the picoseconds to microseconds time scale. *Proc. Natl. Acad. Sci. USA*, 104(7):2163–8.

[Finkelstein and Ptitsyn, 1987] Finkelstein, A. and Ptitsyn, O. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.*, 50(3):171–90.

[Fitzkee et al., 2005] Fitzkee, N., Fleming, P., Gong, H., Panasik, N., Street, T., and Rose, G. (2005). Are proteins made from a limited parts list? *Trends in Biochemical Sciences*, 30(2):73–80.

[Fleming et al., 2006] Fleming, P. J., Gong, H., and Rose, G. D. (2006). Secondary structure determines protein topology. *Prot. Sci.*, 15(8):1829–34.

[Flory, 1969] Flory, P. J. (1969). *Statistical mechanics of chain molecules*. Interscience, New York.

[Flory, 1974] Flory, P. J. (1974). Foundations of rotational isomeric state theory and general methods for generating configurational averages. *Macromolecules*, 7:381–392.

[Frauenfelder, 2002] Frauenfelder, H. (2002). Proteins: Paradigms of complexity. *Proc. Natl. Acad. Sci. USA*, 99:2479–2480.

[Frauenfelder and Leeson, 1998] Frauenfelder, H. and Leeson, D. T. (1998). The energy landscape in non-biological and biological molecules. *Nature Struct. Biol.*, 5(9):757–759.

[Frenkel, 1999] Frenkel, D. (1999). Entropy-driven phase transitions. *Physica A*, 263(1):26–38.

[Frick et al., 1994] Frick, A., Ludwig, A., and Mehldau, H. (1994). A Fast Adaptive Layout Algorithm for Undirected Graphs. *Proceedings of Graph Drawing'94*, 896:388–403.

[Gallagher et al., 1994] Gallagher, T., Alexander, P., Bryan, P., and Gilliland, G. L. (1994). Two crystal structures of the b1 immunoglobulin-binding domain of streptococcal protein g and comparison with nmr. *Biochemistry*, 33(15):4721–4729.

[Gavin et al., 2002] Gavin, A., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A., Cruciat, C., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147.

[Gfeller et al., 2007] Gfeller, D., De Los Rios, P., Caflisch, A., and Rao, F. (2007). Complex network analysis of free-energy landscapes. *Proc. Natl. Acad. Sci. USA*, 104(6):1817–1822.

[Gnanakaran et al., 2003] Gnanakaran, S., Nymeyer, H., Portman, J., Sanbonmatsu, K. Y., and García, A. E. (2003). Peptide folding simulations. *Curr. Opin. Struct. Biol.*, 13(2):168–174.

[Gnedenko, 1954] Gnedenko, B. (1954). *Probability Theory*. Russian Gostekhizdat Moscow.

[Go, 1983] Go, N. (1983). Theoretical studies of protein folding. *Annu Rev Biophys Bioeng*, 12:183–210.

[Gsponer and Caflisch, 2001] Gsponer, J. and Caflisch, A. (2001). Role of native topology investigated by multiple unfolding simulations of four SH3 domains. *J. Mol. Biol.*, 309:285–298.

[Gsponer and Caflisch, 2002] Gsponer, J. and Caflisch, A. (2002). Molecular dynamics simulations of protein folding from transition state. *Proc. Natl. Acad. Sci. USA*, 99:6719–6724.

[Gsponer et al., 2003] Gsponer, J., Haberthur, U., and Caflisch, A. (2003). The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion sup35. *Proc. Natl. Acad. Sci. USA*, 100(9):5154–5159.

[Gutin et al., 1996] Gutin, A. M., Abkevich, V. I., and Shakhnovich, E. I. (1996). Chain length scaling of protein folding time. *Phys. Rev. Lett.*, 77:5433–5436.

[Hänggi and Talkner, 1985] Hänggi, P. and Talkner, P. (1985). First-passage time problems for non-Markovian processes. *Phys. Rev. A*, 32:1934–1938.

[Hänggi et al., 1990] Hänggi, P., Talkner, P., and Borkovec, M. (1990). Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.*, 62:251.

[Haran, 2003] Haran, G. (2003). Single-molecule fluorescence spectroscopy of biomolecular folding. *J. Phys.: Condens. Matter*, 15:R1291–R1317.

[Henry and Eaton, 2004] Henry, E. and Eaton, W. (2004). Combinatorial modeling of protein folding kinetics: free energy profiles and rates. *Chemical Physics*, 307(2-3):163–185.

[Hiltpold et al., 2000] Hiltpold, A., Ferrara, P., Gsponer, J., and Caflish, A. (2000). Free energy calculation of the helical peptide Y(MEARA)6. *J. Phys. Chem. B*, 104:10080–10086.

[Hockenmaier et al., 2007] Hockenmaier, J., Joshi, A. K., and Dill, K. A. (2007). Routes are trees: The parsing perspective on protein folding. *Proteins: Structure, Function, and Bioinformatics*, 66:1–15.

[Hollośi et al., 1985] Hollośi, M., Kawai, M., and Fasman, G. D. (1985). Studies on proline-containing tetrapeptide models of beta-turns. *Biopolymers*, 24(1):211–42.

[Huang, 1987] Huang, K. (1987). *Statistical mechanics*. Wiley New York.

[Ihalainen et al., 2007] Ihalainen, J. A., Bredenbeck, J., Pfister, R., Helbing, J., Chi, L., van Stokkum, I. H. M., Woolley, G. A., and Hamm, P. (2007). Folding and unfolding of a photoswitchable peptide from picoseconds to microseconds. *Proc Natl Acad Sci U S A*, 104(13):5383–5388.

[Itzhaki et al., 1995] Itzhaki, L., Otzen, D., and Fersht, A. (1995). The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods - evidence for a nucleation-condensation mechanism for protein-folding. *J Mol Biol*, 254:260–288.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

[Johnson, 1967] Johnson, S. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

[Jurka and Smith, 1987a] Jurka, J. and Smith, T. F. (1987a). Beta-turn-driven early evolution: the genetic code and biosynthetic pathways. *J Mol Evol*, 25(1):15–9.

[Jurka and Smith, 1987b] Jurka, J. and Smith, T. F. (1987b). Beta turns in early evolution: chirality, genetic code, and biosynthetic pathways. *Cold Spring Harb Symp Quant Biol*, 52:407–10.

[Kabsch and Sander, 1983] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.

[Karanicolas and Brooks, 2003] Karanicolas, J. and Brooks, C. L. (2003). Improved Go-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.*, 334(2):309–325.

[Karpen et al., 1993] Karpen, M. E., Tobias, D. J., and Brooks, C. L. (1993). Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry*, 32(2):412–420.

[Karplus et al., 1987] Karplus, M., Ichiye, T., and Pettitt, B. M. (1987). Configurational entropy of native proteins. *Biophys J*, 52(6):1083–1085.

[Karplus and Weaver, 1979] Karplus, M. and Weaver, D. (1979). Diffusion-collision model for protein folding. *Biopolymers*, 18:1421–1437.

[Karplus and Weaver, 1994] Karplus, M. and Weaver, D. (1994). Protein folding dynamics: The diffusion-collision model and experimental data. *Prot. Sci.*, 3:650–668.

[Kemeny and Snell, 1976] Kemeny, J. and Snell, J. (1976). *Finite Markov Chains*. Springer.

[Khinchin, 1957] Khinchin, A. I. (1957). *Mathematical foundations of information theory*. Dover, New York.

[Kiefhaber et al., 1997] Kiefhaber, T., Bachmann, A., Wildegger, G., and Wagner, C. (1997). Direct measurement of nucleation and growth rates in lysozyme folding. *Biochemistry-Us*, 36:5108–5112.

[Kim and Baldwin, 1982] Kim, P. S. and Baldwin, R. L. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem*, 51:459–489.

[Kim and Baldwin, 1990] Kim, P. S. and Baldwin, R. L. (1990). Intermediates in the folding reactions of small proteins. *Annu Rev Biochem*, 59:631–660.

[Klimov and Thirumalai, 2000] Klimov, D. and Thirumalai, D. (2000). Mechanisms and kinetics of beta-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 97:2544–2549.

[Klimov and Thirumalai, 1996] Klimov, D. K. and Thirumalai, D. (1996). Criterion that determines the foldability of proteins. *Phys. Rev. Lett.*, 76(21):4070–4073.

[Kornai, 2002] Kornai, A. (2002). How many words are there? *Glottometrics*, 4:61–86.

[Koshiba et al., 2001] Koshiba, T., Kobashigawa, Y., Demura, M., and Nitta, K. (2001). Energetics of three-state unfolding of a protein: canine milk lysozyme. *Protein Eng*, 14(12):967–974.

[Krivov and Karplus, 2002] Krivov, S. and Karplus, M. (2002). Free energy disconnectivity graphs: Application to peptide models. *The Journal of Chemical Physics*, 117:10894.

[Kuhlman et al., 2003] Kuhlman, B., Dantas, G., Ireton, G., Varani, G., Stoddard, B., and Baker, D. (2003). Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302(5649):1364–1368.

[Kuwajima, 1989] Kuwajima, K. (1989). The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins: Structure, Function, and Bioinformatics*, 6:87–103.

[Kyte and Doolittle, 1982] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

[Landau and Lifshitz, 1980] Landau, L. D. and Lifshitz, E. M. (1980). *Statistical Physics*. Pergamon, Oxford.

[Landsberg, 1984] Landsberg, P. T. (1984). Can entropy and "order" increase together? *Phys. Lett.*, 102A:171–173.

[Larson et al., 2002] Larson, S., Snow, C., Shirts, M., and Pande, V. (2002). Folding@home and genome@home: Using distributed computing to tackle previously intractable problems in computational biology. *Computational Genomics*.

[Lee et al., 2003] Lee, C. L., Stell, G., and Wang, J. (2003). First-passage time distribution and non-Markovian diffusion dynamics of protein folding. *J. Chem. Phys.*, 118:959–968.

[Lenz et al., 2004] Lenz, P., Zagrovic, B., Shapiro, J., and Pande, V. S. (2004). Folding probabilities: A novel approach to folding transitions and the two-dimensional Ising model. *J. Chem. Phys.*, 120:6769–6778.

[Levinthal, 1968] Levinthal, C. (1968). Are there pathways for protein folding? *Journal de Chimie Physique*, 65(1):44–45.

[Li et al., 2008] Li, C., Yang, H., and Komatsuzaki, T. (2008). Multiscale complex network of protein conformational fluctuations in single-molecule time series. *Proc. Natl. Acad. Sci. USA*.

[Lieberman et al., 2007] Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., and Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–6.

[Lipman et al., 2003] Lipman, E. A., Schuler, B., Bakajin, O., and Eaton, W. A. (2003). Single-molecule measurement of protein folding kinetics. *Science*, 301(5637):1233–1235.

[Ma and Dinner, 2005] Ma, A. and Dinner, A. R. (2005). Automatic method for identifying reaction coordinates in complex systems. *J Phys Chem B*, 109(14):6769–6779.

[Makhatadze and Privalov, 1995] Makhatadze, G. I. and Privalov, P. L. (1995). Energetics of protein structure. *Adv Protein Chem*, 47:307–425.

[Mayor et al., 2003] Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J. G., Sato, S., Jas, G. S., Freund, S. M. V., Alonso, D. O. V., Daggett, V., and Fersht, A. R. (2003). The complete folding pathway of a protein from nanoseconds to microseconds. *Nature*, 421(6925):863–867.

[Mirsky and Pauling, 1936] Mirsky, A. E. and Pauling, L. (1936). On the structure of native, denatured, and coagulated proteins. *Proc. Natl. Acad. Sci. USA*, 22(7):439–447.

[Müller et al., 1997] Müller, R., Talkner, P., and Reimann, P. (1997). Rates and mean first passage times. *Physica A*, 247:338–356.

[Muñoz, 2001] Muñoz, V. (2001). What can we learn about protein folding from Ising-like models? *Curr. Opin. Struct. Biol.*, 11(2):212–216.

[Muñoz et al., 1997] Muñoz, V., Thompson, P., Hofrichter, J., and Eaton, W. (1997). Folding dynamics and mechanism of beta-hairpin formation. *Nature*, 390:196–199.

[Nerukh et al., 2004] Nerukh, D., Karvounis, G., and Glen, R. C. (2004). Quantifying the complexity of chaos in multibasin multidimensional dynamics of molecular systems. *Complexity*, 10:40–46.

[Neuweiler and Sauer, 2004] Neuweiler, H. and Sauer, M. (2004). Using photoinduced charge transfer reactions to study conformational dynamics of biopolymers at the single-molecule level. *Curr Pharm Biotechnol*, 5(3):285–298.

[Noe et al., 2007] Noe, F., Horenko, I., Schutte, C., and Smith, J. C. (2007). Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J Chem Phys*, 126(15):155102.

[Nolting and Andert, 2000] Nolting, B. and Andert, K. (2000). Mechanism of protein folding. *Proteins: Structure, Function, and Genetics*, 41:288–298.

[Olivieri and Scoppola, 1996] Olivieri, E. and Scoppola, E. (1996). Markov chains with exponentially small transition probabilities: First exit problem from a general domain. II. The general case. *Journal of Statistical Physics*, 84(5):987–1041.

[Onuchic et al., 1997] Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600.

[Ozkan et al., 2001] Ozkan, S. B., Bahar, I., and Dill, K. A. (2001). Transition states and the meaning of Phi-values in protein folding kinetics. *Nature Struct. Biol.*, 8:765–769.

[Ozkan et al., 2003] Ozkan, S. B., Dill, K. A., and Bahar, I. (2003). Computing the transition state populations in simple protein models. *Biopolymers*, 68:35–46.

[Ozkan et al., 2007] Ozkan, S. B., Wu, G. A., Chodera, J. D., and Dill, K. A. (2007). Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. USA*, 104(29):11987–92.

[Paci et al., 2003] Paci, E., Cavalli, A., Vendruscolo, M., and Caflisch, A. (2003). Analysis of the distributed computing approach applied to the folding of a small beta peptide. *Proc. Natl. Acad. Sci. USA*, 100(14):8217–8222.

[Pande et al., 2000] Pande, V. S., Grosberg, A. Y., and Tanaka, T. (2000). Heteropolymer freezing and design: towards physical models of protein folding. *Rev. Mod. Phys.*, 72:259–314.

[Pappu et al., 2000] Pappu, R. V., Srinivasan, R., and Rose, G. D. (2000). The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl. Acad. Sci. USA*, 97(23):12565–12570.

[Park and Levitt, 1995] Park, B. and Levitt, M. (1995). The complexity and accuracy of discrete state models of protein-structure. *J. Mol. Biol.*, 249:493–507.

[Park and Pande, 2006] Park, S. and Pande, V. S. (2006). Validation of Markov state models using Shannon's entropy. *J. Chem. Phys.*, 124:054118.

[Pellarin and Caflisch, 2006] Pellarin, R. and Caflisch, A. (2006). Interpreting the aggregation kinetics of amyloid peptides. *J. Mol. Biol.*, 360(4):882–892.

[Pellarin et al., 2007] Pellarin, R., Guarnera, E., and Caflisch, A. (2007). Pathways and intermediates of amyloid fibril formation. *J. Mol. Biol.*, 379:917–924.

[Peters et al., 2007] Peters, B., Beckham, G. T., and Trout, B. L. (2007). Extensions to the likelihood maximization approach for finding reaction coordinates. *J Chem Phys*, 127(3):034109.

[Plaxco et al., 1998] Plaxco, K., Simons, K., and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, 277:985–994.

[Plotkin and Wolynes, 1998] Plotkin, S. S. and Wolynes, P. G. (1998). Non-Markovian configurational diffusion and reaction coordinates for protein folding. *Phys. Rev. Lett.*, 80:5015–5018.

[Privalov and Khechinashvili, 1974] Privalov, P. L. and Khechinashvili, N. N. (1974). A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J Mol Biol*, 86(3):665–684.

[Ptitsyn, 1992] Ptitsyn, O. (1992). *Protein Folding: The molten globule state*. W. H. Freeman and Company.

[Ptitsyn and Finkelstein, 1980] Ptitsyn, O. and Finkelstein, A. (1980). Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? *Q Rev Biophys*, 13(3):339–86.

[Ptitsyn and Rashin, 1975] Ptitsyn, O. and Rashin, A. (1975). Model of myoglobin self-organization. *Biophys Chem*, 3:1–20.

[Ptitsyn and Uversky, 1994] Ptitsyn, O. and Uversky, V. (1994). The molten globule is a third thermodynamical state of protein molecules. *FEBS Lett*, 341(1):15–8.

[Ptitsyn, 1995] Ptitsyn, O. B. (1995). Molten globule and protein folding. *Adv Protein Chem*, 47:83–229.

[Ptitsyn et al., 1990] Ptitsyn, O. B., Pain, R. H., Semisotnov, G. V., Zerovnik, E., and Razgulyaev, O. I. (1990). Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett*, 262(1):20–24.

[Rackovsky and Scheraga, 1977] Rackovsky, S. and Scheraga, H. (1977). Hydrophobicity, hydrophilicity, and radial and orientational distributions of residues in native proteins. *Proc. Natl. Acad. Sci. USA*, 74:5248–5251.

[Ramachandran et al., 1963] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–99.

[Ramirez-Alvarado et al., 2000] Ramirez-Alvarado, M., Merkel, J., and Regan, L. (2000). A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. *Proc. Natl. Acad. Sci. USA*, 97:8979–8984.

[Ramirez-Alvarado and Regan, 2002] Ramirez-Alvarado, M. and Regan, L. (2002). Does the location of a mutation determine the ability to form amyloid fibrils? *J. Mol. Biol.*, 323:17–22.

[Rao and Caflisch, 2003] Rao, F. and Caflisch, A. (2003). Replica exchange molecular dynamics simulations of reversible folding. *The Journal of Chemical Physics*, 119:4035.

[Rao and Caflisch, 2004] Rao, F. and Caflisch, A. (2004). The protein folding network. *J. Mol. Biol.*, 342(1):299–306.

[Rao et al., 2005] Rao, F., Settanni, G., Guarnera, E., and Caflisch, A. (2005). Estimation of protein folding probability from equilibrium simulations. *J. Chem. Phys.*, 122(18):184901.

[Regan and De Grado, 1988] Regan, L. and De Grado, W. (1988). Characterization of a helical protein designed from first principles. *Science*, 241(4868):976–978.

[Riddle et al., 1997] Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q., and Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.*, 4(10):805–809.

[Robinson and Clark, 1999] Robinson, C. and Clark, R. (1999). *Dynamical Systems: stability, symbolic dynamics, and chaos*. CRC Press.

[Rose et al., 2006] Rose, G. D., Fleming, P. J., Banavar, J. R., and Maritan, A. (2006). A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci. USA*, 103(45):16623–16633.

[Rose et al., 1985] Rose, G. D., Gierasch, L. M., and Smith, J. A. (1985). Turns in peptides and proteins. *Advan. Protein Chem.*, 37:1–109.

[Rylance et al., 2006] Rylance, G. J., Johnston, R. L., Matsunaga, Y., Li, C.-B., Baba, A., and Komatsuzaki, T. (2006). Topographical complexity of multidimensional energy landscapes. *Proc. Natl. Acad. Sci. USA*, 103(49):18551–5.

[Sabelko et al., 1999] Sabelko, J., Ervin, J., and Gruebele, M. (1999). Observation of strange kinetics in protein folding. *Proc Natl Acad Sci U S A*, 96(11):6031–6036.

[Sánchez and Kiefhaber, 2003] Sánchez, I. and Kiefhaber, T. (2003). Evidence for Sequential Barriers and Obligatory Intermediates in Apparent Two-state Protein Folding. *Journal of Molecular Biology*, 325(2):367–376.

[Schaefer et al., 2006] Schaefer, M., Guarnera, E., and Paci, E. (2006). Coarse grained descriptions and quantitative analysis in computational protein folding. *Unpublished*.

[Schaefer et al., 1998] Schaefer, M., van Vlijmen, H. W. T., and Karplus, M. (1998). Electrostatic contributions to molecular free energies in solution. *Advan. Protein Chem.*, 51:1–57.

[Schuler and Eaton, 2008] Schuler, B. and Eaton, W. A. (2008). Protein folding studied by single-molecule fret. *Curr. Opin. Struct. Biol*, In press.

[Schuler et al., 2002] Schuler, B., Lipman, E. A., and Eaton, W. A. (2002). Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419(6908):743–747.

[Scoppola, 1993] Scoppola, E. (1993). Renormalization group for Markov chains and application to metastability. *Journal of Statistical Physics*, 73(1):83–121.

[Shakhnovich, 2006] Shakhnovich, E. (2006). Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev.*, 106(5):1559–1588.

[Shakhnovich et al., 1996] Shakhnovich, E., Abkevich, V., and Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature*, 379:96–98.

[Shakhnovich, 1998] Shakhnovich, E. I. (1998). Protein design: a perspective from simple tractable models. *Folding & Design*, 3(3):R45–R58.

[Shalizi and Crutchfield, 2002] Shalizi, C. and Crutchfield, J. (2002). Information bottlenecks, causal states, and statistical relevance bases: How to represent relevant information in memoryless transduction. *Advances in Complex Systems*, 5(1):91–95.

[Shalizi et al., 2002] Shalizi, K., Shalizi, C., and Crutchfield, J. (2002). Pattern Discovery in Time Series, part I and II: Implementation, Evaluation, and Comparison. *Journal of Machine Learning Research*, pages 02–10.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of comunication. *AT&T Tech. J.*, 27:379–423.

[Shapiro and Varian, 1999] Shapiro, C. and Varian, H. (1999). *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press.

[Shea and Brooks, 2001] Shea, J. E. and Brooks, C. L. r. (2001). From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu Rev Phys Chem*, 52:499–535.

[Shiner et al., 1999] Shiner, J. S., Davison, M., and Landsberg, P. T. (1999). Simple measure for complexity. *Phys. Rev. E*, 59:1459–1464.

[Shortle, 1996] Shortle, D. (1996). The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.*, 10(1):27–34.

[Shortle and Ackerman, 2001] Shortle, D. and Ackerman, M. S. (2001). Persistence of native-like topology in a denatured protein in 8 M urea. *Science*, 293(5529):487–489.

[Simon, 1962] Simon, H. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482.

[Snell, 1959] Snell, J. (1959). Finite Markov Chains and their Applications. *The American Mathematical Monthly*, 66(2):99–104.

[Strait and Dewey, 1996] Strait, B. J. and Dewey, T. G. (1996). The shannon information entropy of protein sequences. *Biophys. J.*, 71(1):148–155.

[Swope et al., 2004a] Swope, W. C., Pitera, J. W., and Suits, F. (2004a). Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J. Phys. Chem. B*, 108:6571–6581.

[Swope et al., 2004b] Swope, W. C., Pitera, J. W., Suits, F., and et al. (2004b). Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and beta-hairpin peptide. *J. Phys. Chem. B*, 108:6582–6594.

[Szabo et al., 1980] Szabo, A., Schulten, K., and Schulten, Z. (1980). First passage time approach to diffusion controlled reactions. *J. Chem. Phys.*, 72:4350–4357.

[Talaga et al., 2000] Talaga, D. S., Lau, W. L., Roder, H., Tang, J., Jia, Y., DeGrado, W. F., and Hochstrasser, R. M. (2000). Dynamics and folding of single two-stranded coiled-coil peptides studied by fluorescent energy transfer confocal microscopy. *Proc Natl Acad Sci U S A*, 97(24):13021–13026.

[Taverna and Goldstein, 2002] Taverna, D. M. and Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins*, 46(1):105–109.

[Thirumalai et al., 2002] Thirumalai, D., Klimov, D. K., and I., D. R. (2002). Insights into specific problems in protein folding using simple concepts. *Advan. Chem. Phys.*, 120:35–76.

[Tompa, 2002] Tompa, P. (2002). Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533.

[Uversky, 2002] Uversky, V. N. (2002). What does it mean to be natively unfolded? *Eur J Biochem*, 269(1):2–12.

[van Gunsteren et al., 2001] van Gunsteren, W. F., Bürgi, R., Peter, C., and Daura, X. (2001). The key to solving the protein-folding problem lies in an accurate description of the denatured state. *Angew Chem Int Ed Engl*, 40(2):351–355.

[van Kampen, 1981] van Kampen, N. G. (1981). *Stochastic processes in physics and chemistry*. North Holland.

[Vassilenko and Uversky, 2002] Vassilenko, K. S. and Uversky, V. N. (2002). Native-like secondary structure of molten globules. *Biochim Biophys Acta*, 1594(1):168–177.

[Vendruscolo et al., 2001] Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409(6820):641–645.

[Voelz and Dill, 2007] Voelz, V. A. and Dill, K. A. (2007). Exploring zipping and assembly as a protein folding principle. *Proteins: Structure, Function, and Bioinformatics*, 66(4):877–88.

[Voronoi, 1908] Voronoi, G. F. (1908). Nouvelles applications des parametres continusa la theorie des formes quadratiques. *Z. Reine Angew. Math.*, 134:198–287.

[Wagner and Kiefhaber, 1999] Wagner, C. and Kiefhaber, T. (1999). Intermediates can accelerate protein folding. *Proc. Natl. Acad. Sci. USA*, 96(12):6716–6721.

[Wales et al., 1998] Wales, D. J., Miller, M. A., and Walsh, T. R. (1998). Archetypal energy landscape. *Nature*, 334:758–760.

[Wang and Wang, 1999] Wang, J. and Wang, W. (1999). A computational approach to simplifying the protein folding alphabet. *Nature Struct. Biol.*, 6(11):1033–1038.

[Watters and Baker, 2004] Watters, A. L. and Baker, D. (2004). Searching for folded proteins in vitro and in silico. *Eur J Biochem*, 271(9):1615–1622.

[Watters et al., 2007] Watters, A. L., Deka, P., Corrent, C., Callender, D., Varani, G., Sosnick, T., and Baker, D. (2007). The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell*, 128(3):613–624.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[Wilmot and Thornton, 1988] Wilmot, C. and Thornton, J. (1988). Analysis and prediction of the different types of beta-turn in proteins. *J Mol Biol*, 203(1):221–32.

[Wolynes, 1997] Wolynes, P. G. (1997). As simple as can be? *Nature Struct. Biol.*, 4(11):871–874.

[Wong, 1975] Wong, J. (1975). A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA*.

[Wu, 1931] Wu, H. (1931). Studies on denaturation of proteins. xiii. a theory of denaturation. 1931. *Chinese J. Physiol.*, 1:219–234.

[Yang et al., 2003] Yang, H., Luo, G., Karnchanaphanurach, P., Louie, T.-M., Rech, I., Cova, S., Xun, L., and Xie, X. S. (2003). Protein conformational dynamics probed by single-molecule electron transfer. *Science*, 302(5643):262–266.

[Yook et al., 2002] Yook, S.-H., Jeong, H., and Barabasi, A.-L. (2002). Modeling the internet's large-scale topology. *Proc. Natl. Acad. Sci. USA*, 99(21):13382–13386.

[Zimm and Bragg, 1958] Zimm, B. H. and Bragg, J. K. (1958). Theory of the one-dimensional phase transition in polypeptide chains. *J. Chem. Phys.*, 28:1246–1247.

[Zimm and Bragg, 1959] Zimm, B. H. and Bragg, J. K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.*, 31:526–535.

[Zwanzig, 1995] Zwanzig, R. (1995). Simple model of protein kinetics. *Proc. Natl. Acad. Sci. USA*, 92:9801–9804.

[Zwanzig, 1997] Zwanzig, R. (1997). Two-state models of protein folding kinetics. *Proc. Natl. Acad. Sci. USA*, 94(1):148–50.

[Zwanzig et al., 1992] Zwanzig, R., Szabo, A., and Bagchi, B. (1992). Levinthal's paradox. *Proc. Natl. Acad. Sci. USA*, 89(1):20–22.

List of Figures

1.1 Possible mechanism for protein folding. (A) The hierarchical model [Ptitsyn and Rashin, 1975, Kim and Baldwin, 1982, Kim and Baldwin, 1990, Baldwin and Rose, 1999b, Baldwin and Rose, 1999a, Rose et al., 2006, Ozkan et al., 2007]. Protein folding is thought to start with the formation of elements of secondary structure independently of tertiary structure, or at least before tertiary structure is locked in place. These elements then assemble into the tightly packed native tertiary structure by means of a diffusion-collision (or framework) mechanism [Karplus and Weaver, 1979, Karplus and Weaver, 1994]. (B) Hydrophobic collapse model for folding [Rackovsky and Scheraga, 1977, Dill, 1985, Dill, 1990b]. The initial event of the reaction is thought to be a relatively uniform collapse of the protein molecule, mainly driven by a phase separation given by the hydrophobic effect. Stable secondary structure starts to grow only from the collapsed state. (C) Nucleation-condensation mechanism [Fersht, 1995, Itzhaki et al., 1995, Shakhnovich et al., 1996, Fersht, 1997, Fersht, 1999, Kiefhaber et al., 1997]. Early formation of a diffuse protein-folding nucleus catalyses further folding. The nucleus primarily consists of a few adjacent residues which have some correct secondary structure interactions, but it is stable only in the presence of further approximately correct tertiary structure interactions. Both mechanisms (B) and (C) at final stage of folding are compatible with the funnel model [Bryngelson et al., 1995, Dill and Chan, 1997]. Figure adapted from [Nolting and Andert, 2000]. 5

2.1 The distribution of the pairwise Cα-RMSD computed from an equilibrium simulation of the GSGS peptide. 19

2.2 Ensemble Ramachandran 2D histogram with the partitions used to construct the strings of rotational states for a polypeptide chain SRA[4]. The 2D histogram is obtained from the cumulative frequencies of all (ϕ, ψ) pairs on a 360×360 grid. Frequencies are estimated from the GSGS MD trajectory. 22

2.3 The distribution of the total energy (the sum of potential and solvation energies) from the whole ensemble of microstates of the GSGS simulation. The agreement with a gaussian is excellent, with mean effective energy $\overline{E} = -4.04$ kcal/mol and standard deviation $\sigma(E) = 10.74$ kcal/mol. 27

2.4 The energy (left column) and the free energy (right column) distributions among the ensemble of mesostates: black points are weighted with the mesostate probabilities while the red ones are not. Peaks in the weighted energy distributions corresponds to the folded mesostate. The unweighted energy distributions follows always a gaussian. The weighted free energy distributions are double peaked corresponding respectively to the folded and the unfolded state. 28

2.5 (A) The sublinear increasing of the number of visited mesostates N as a function of the number of sampled microstates M for different coarse graining procedures; (B) the ratio N/M which is the probability to sample a new mesostate over the time. 36

2.6 The density of mesostates of the GSGS from the descriptions based on strings (A) and RMSD clustering (B). 37

2.7 An example to explain how clustering based on first neighboring produces spurious unassigned microstates. 38

2.8 The ranking distributions for the GSGS from the descriptions based on strings (A) and RMSD clustering (B). 41

2.9 Selected mesostates corresponding to the description based on SRA[4] taken among the first $r_c = 262$ mesostates. The figures represent ensemble of structures with their fluctuations within a mesostate corresponding to a well defined string. The N-term of the polypeptide is colored in red. The structures are represented with their RMSF fluctuations by using the macro "sausage" implemented in program molmol. 42

2.10 The string site probabilities for the mesostates based on SSS[4] (left) and SRA[4] (right) of the GSGS. The similarities between the two descriptions are evident. 44

2.11 The distributions of the combined probabilities for SRA[4] in a log-log plot. The black dots are the observed string probabilities as a function of the combined probabilities while the red curve is the density of combined probabilities. The latter fits very well with triple log-normal distribution (dashed curve) which suggests that the ensemble of strings is organized in three sub-ensembles. The dotted curves are single log-normal distributions with the parameters estimated from the triple fit. 45

2.12 (A) Disorder per residue for the string based on secondary structure SSS[8], SSS[4] and on rotational angles SRA[4]; (B) the map of native contacts whose points represent the disorder of a Ca contact on which the strings SNC[2] are based. 46

2.13 (A) SRA[4]; (B) SSS[8]. The configurational hierarchy of the folded string. Heterogeneity in the maps implies the existence of patterns in the configurational space. Patterns are revealed in terms of hierarchical trees which are constructed combining all the contiguous folded substrings in such a way to obtain fragments of the folded string that have maximal probability. For instance, at the lowest hierarchical level we have R 1-fragments which correspond to R nodes on the maps. At the next hierarchical level the 1-fragments are combined to obtain 2-fragments: two contiguous 1-fragments are combined if the probability of the corresponding 2-fragment is higher than the alternative 2-fragments. The 2-fragment so obtained gets a node which is linked to previous nodes of the 1-fragments, and so on. If a 1-fragment does not combine with any 2-fragment then it can be combined at next hierarchical level to form fragments of length longer than 2. . 48

2.14 Disorder $\mathcal{D}(H)$ (A) and $\mathcal{D}(h)$ (B) as a function of the number of sampled microstates M. The convergence of the simulations is a signature of statistical convergence. This arises from the compensation between folding and unfolding events: former increase disorder while the latter decrease it. 50

2.15 The Cα-RMSD distribution with respect to the folded structure of the GSGS from which a FPT distribution for unfolding is estimated. The unfolded target state has been chosen with Cα-RMSD greater than 6 Å. 51
2.16 First passage time distributions to the folded mesostate for all the descriptions adopted to study the GSGS. Double and triple exponential functions were used to fit the data. . . . 53
2.17 The mean first passage time MFPT distributions to the folded mesostate for the adopted descriptions. The distributions show a clear kinetic partitioning at all time scales in both the folded and unfolded phases: the pronounced peaks in the main unfolded region correspond to mesostates having well defined relaxation times to the folded state. 55
2.18 The relation between thermodynamics and kinetics by using the MFPT as a reaction coordinate and SRA[4] as coarse graining. 56
2.19 The kinetic hierarchies of the folded string for the descriptions SRA[4] (A) and SSS[8] (B). A pattern of folding pathways appears in figure **??** as well. The trees composed by black nodes and edges are extrapolated from the maps by using the same algorithm employed to construct the trees of figure **??**. For instance, two contiguous 1-fragments (two nodes at the lowest hierarchy) are assembled together to form a 2-fragment (giving a new node at the next hierarchical level) if the formation rate of it is the highest possible among all the possible 2-fragments that 1-fragments can form. The trees show the fastest way in which the folded string can be assembled from all folded substrings. 58
2.20 (A) The local rates for the mesoscopic transitions k_0 ($c \to i$) and k_1 ($i \to c$) that are used for the generalized Zwanzig model according to the SRA[4] description. (B) The ratio ϵ/k_BT corresponding to the favorable energy bias to a correct bond chain from equation **??**. . . . 60
2.21 The procedure to construct the master equation on the s quantity. 61
2.22 The amount of non Markov fluxes as a function of the number of steps at which the transition matrix T is estimated for 4 kind of mesoscopic descriptions of the configurational space of the GSGS. 66
2.23 The idea behind the causal grouping algorithm. 67
2.24 The convergence of the sampled number of causal grouped mesostates as a function of the length of the input simulation with a cutoff of 200 ($6 \cdot 10^5$ microstates corresponds to 12 μs of simulation time). 68
2.25 The non-Markov fluxes as a function of the number of steps at which the transition matrix T is estimated for the mesoscopic descriptions SRA and SSS respectively for a causal grouping at 200, 300 and 1000 mesostates in comparison with the non-Markov fluxes of the ungrouped descriptions SNC[2], SSS[8], SRA[4] and RMSD[1.5]. 69
2.26 The relaxation time scales of the Markov chains computed from the inverse of the rate spectrum of the matrix $K = \mathbf{1} - T$ estimated on the causal grouped time series SRA[4] and SSS[8]. 71
2.27 The normalized eigenvector corresponding to the slowest non null rate k_2 of the Markov chains constructed on the causal mesostates SRA[4] (A) and SSS[8] (B). Some intensity peaks are shown together with the ensemble of structures they correspond to. Pictures have been made with the program molmol. 72

2.28 The network corresponding to the transition matrix estimated from 1000 causal grouped mesostates based on the SRA[4]. The total number of vertices are 1000 and the edges are 15608 without any cutoff on the probabilities. The graph has been realized with Tulip and the edges have been coloured according to the value they assume on the transition matrix. It appears clear that two main phases characterize the transition matrix: a helix phase clearly separated from a beta phase at whose center is posed the folded state. . . . 74

2.29 The network corresponding to the transition matrix estimated from 1000 causal grouped mesostates based on the SSS[4]. The total number of vertices are 1000 and the edges are 33760 without any cutoff on the probabilities. The graph has been realized with Tulip and the edges have been coloured according to the value they assume on the transition matrix. 75

2.30 The transition matrix graphs corresponding to 200 causal mesostates for SRA[4] (top) and SSS[8] (bottom). Graphs have been realized with the program Tulip. Circles and boxes represent basins clearly distinguishable from other vertices. Indicative populations of the basins are reported. In the bottom network a helical basin is missing, helices are present as transient nodes (see the node 99 in the bottom network) as the rank of the first helix SSS[8] mesostate is much lower than that in SRA[4] (the 99th "`---HHHHHHHHHS------`" for SSS[8] and 40th "`0100011111111111111`" for SRA[4]) 76

2.31 The inverse of the recurrence time $1/M^*_{i\to i}$ (red curve) referred to the left y axis; the black curve (right y axes) represents the MFPT to the folded state on the reordered MFPT matrix, namely its first row $M^*_{i\to 1}$. The x axes gives the reordered mesostate id from low MFPT to large MFPT to the folded state. Averaged structures are shown over representatives peaks of the recurrence time. Causal mesostates from SRA[4]. 78

2.32 The reordered MFPT matrix $M^*_{i\to j}$ shows that the native basin works as a "hub" in the overall kinetics. An entry on the matrix gives the MFPT for the equilibrium transition $i \to j$ from causal grouped mesostates SRA[4]. Horizontal bands are equilibrium transitions from all the is to a specific j. Yellow points can be assumed to have an infinite barrier between initial and final state so that their inter-conversion rate is zero. Helix-like and curl-like do not exchange between them in a time scale that is comparable with their folding time. They directly relax to to folded state in about 90 ns and 60 ns respectively. . 79

2.33 The free energy landscape emerging from the Markovian treatment of the GSGS kinetics. Low entropy states (curl-like states) act as kinetic traps while the helical-phase represents the unfolded state that is stabilized by high entropy. The folded basin is stabilized by both enthalpy and entropy. Folding can then initiate either from low enthalpy or high entropy states. 81

3.1 Amino acid propensities (left y axis and black curve) from the dataset found in [Dayalan et al., 2006]. Hydropathy index of the amino acids (right y axes and red bars) from [Kyte and Doolittle, 1982]. 86

3.2 Amino acid informations for secondary structure SSS[8] (empty squared) and mesoscopic dihedral states SRA[4] (gray bars) estimated from the dataset found in [Dayalan et al., 2006] 87

3.3 (A) The ensemble representation of the most populated folded mesostates for the polyTHR_xGS, respectively three-, four-, five-, six-stranded β-sheets. These mesostates are mainly promoted by the low enthalpy. (B) The normalized Shannon entropy per residue (also called disorder) from the ensemble of SRA[4] mesostates obtained from the polyTHR simulations. High disorder peaks correspond to the Gly residues at the turn positions. Disorder profiles look very similar among the sequences showing a modular pattern. (C) The total Shannon entropy (black curve) and the mean disorders per residue respectively as functions of the number of GS turns. The disorder does not depend on the protein size while the total entropy linearly increases with the number of GS turns, that is the chain size. 93

3.4 Time series Cα-RMSD of the polyTHR_xGS folding simulations at 330 K with respect to the averaged β-strand structures. The model mean structures are computed from the ensemble of structures which are members of the most populated SRA[4] mesostate. Many spontaneous folding events are observed for all four the proteins, they result to be about 250 for polyTHR_2GS (A), 110 for polyTHR_3GS (B), 40 for polyTHR_4GS (C) and 20 for polyTHR_5GS (D). 94

3.5 Time series Cα-RMSD of the 1pgb_AGT for the stability simulations at 300 K (A) and folding simulations at 330 K (B) with respect to the X-ray structure 1pgb. The firsts and lasts 2 Cα and the Gly residues were excluded from the RMSD calculations. In the time series at 330 K approximatively 30 folding events can be identified. In the time series at 300 K the folded state in maintained in its topology for about 500 ns and an unfolding event can be assumed as a Poissonian event. (C) The mean effective energy was computed on Cα-RMSD ranges of 0.5 Å for both the simulations performed at 300 and 330 K. On plot (C) the effective mean energy differences $\Delta E_{\mathrm{eff}}(\mathrm{RMSD}) = \langle E_{\mathrm{eff}}(\mathrm{RMSD})\rangle - \langle E_{\mathrm{eff}}(\mathrm{RMSD} \leq 2)\rangle$ as a function of the Cα-RMSD with respect to the X-ray structure. The curves clearly indicates that the protein G topology is the enthalpic minima for protein 1pgb_AGT at both the simulation temperatures. An enthalpy difference can be estimated between folded and unfolded: approximatively 19 kcal/mol at 330 K and 9 kcal/mol at 300 K. In (D) a two dimensional density plot of the Cα-RMSD as a function of the Dihedral-RMSD, both with respect to the X-ray structure 1pgb. All the residues were taken into account. The plot clearly shows the presence of two broad phases, one ranging from 2.5 to 6 Å and from 55 to 70 deg in Dihedral-RMSD and the other much broader. Within the first phase the folded state can be located. Unlike the Cα-RMSD the Dihedral-RMSD is more sensitive to the local similarities to the X-ray structure. Interestingly, in the broader phase the Dihedral-RMSD includes native like values ($\sim$60 deg) suggesting that in the unfolded state the elements of secondary structure can be locally already shaped. 96

3.6 (A left) The ensemble representation of the most populated and folded cluster of structures: the mean pairwise Cα-RMSD within the cluster is 3.5 Å (all the Cα considered). (A right) The Cα-RMSD structural alignment between the folded cluster center (red ribbon) and the X-ray structure (blue ribbon) of protein G (1pgd pdb code), Cα-RMSD between the two structures is 2.6 Å where the first and last 2 Cα and the Gly residues were excluded from the calculation. (B) The Cα-RMSD with respect to the X-ray structure within the folded cluster of structures as a function of the number of Cα pairs used to compute the RMSD. Given a number of Cα pairs the structural alignment finds the best overlap between chain fragments that can also be not contiguous along the sequence. The black circles are mean Cα-RMSD values with their standard deviations while the red diamonds are the best Cα-RMSD values for a given number of Cα pairs. From the 50% up to the 85% of the Cα pairs the corresponded Cα-RMSD steadily turns out lower than 2.6 Å while the best values are around 1.5 Å. The result provides a quantitative indication of the fluctuating nature of the folded cluster and yet shows that the folded topology of protein G can be satisfied in a shallow manner, hairpins for instance can be still formed with a chain shifting up to ± 2 residues. 99

3.7 Protein 1pgb_AGT: (A) the normalized entropies per residue respectively: for the ensembles of SRA[4] string of mesostates corresponding to the RMSD[5.0] folded cluster (black curve), for the whole ensemble of strings (red curve) and for the sub-ensemble of strings such that the corresponding microstates have Cα-RMSD to the X-ray structure greater than 10 Å (bleu curve). The latter sub-ensemble of strings is clearly referred solely to the unfolded state of the protein. It is interesting to notice the slightly difference of the entropy profiles between the red and blue curve. The profile of the folded cluster is very peaked to the loop regions showing their high disorder, on the contrary the first and last strands and the helix regions appear very ordered. In (B) and (C) the statistics of SRA[4] and SSS[4] mesostates per residue are reported for the whole ensemble of strings. The SSS[4] description is based on the DSSP alphabet for secondary structure in which the following grouped states were considered: beta=E+B, helix=H+I, turn/loop=S+T+G, coil. The helix profile of SSS[4] in (C) is compatible with the helix/turn/loop profile SRA[4] in (B) meaning that the two descriptions are very similar. The peaks of disorder in (A) are based on the dihedral description, so that they essentially take into account the disorder due to the coil structures located at the N- and C terminals, and due to the different kind of turn/loop configurations that can be realized with several dihedral arrangements. 101

3.8 On the top the probability of the possible contiguous folded substrings is shown: an entry in the triangular map represents the estimated probability, computed on the time series, of a chain folded fragment having length going from 1 (single residue) to the full length chain. The length of a substring gives the hierarchy level on the y axes of the maps. The fragment probabilities are computed on the whole ensemble of strings discarding the full folded string, that is to reduce the bias on the substring probabilities due to the full folded string population. The maps play the role of free energy landscapes with respect to the progress variable hierarchy length, that is the number of residues that are folded. Highly ordered heterogeneity appears on the maps which can be interpreted as the existence of patterns in the ensemble of strings. From the maps hierarchy trees can be extrapolated as previously shown in chapter **??**. At the lowest hierarchical level R walkers (corresponding to R residues) start a random walk, namely that 1-fragments are assembled in a certain way to gain the next status level of 2-fragments. The algorithm makes the walkers follow the maximal probability route, for instance two 1-fragments can assemble to two different 2-fragments, thus the algorithm shall choose that maximizing the 2-fragment probability. The procedure is repeated for all the hierarchies until a tree is completed by reaching the full folded string that lies on the top of the tree. The algorithm finds the maximal probability tree associated to the map. . 103

3.9 PolyTHR_xGS: in (A) the folding times are shown as a function of the number of turns of the β-stranded folded states. These times scale exponentially with the number of turns of the folded state and give e pre-exponential factor of about 12.8 ns which can be interpreted as the diffusion time of an hairpin (see text). In (B) the energy difference of the folded string (black circles) $\Delta E_{\rm fold} = \overline{E}_{\rm fold} - \overline{E}$ and the configurational entropy loss of the folded string (empty diamonds) $T\Delta h_{\rm fold} = h_{\rm fold} - h$ scaling linearly with the number of turns. . 105

3.10 1pgb_AGT: the FPT distributions for folding (black data) and unfolding (red data). As target state for folding the most populated cluster RMSD[5.0] (see figure **??** (A)) was used while a threshold of 13 Å in the time series of the Cα-RMSD to the X-ray structure was employed to define an unfolded phase. All the distributions fit very well with a double exponential function in which the slow phase corresponds to folding (unfolding) and the fast phase represents a diffusion within the folded state. From the fits it turns out $t_{\rm fold} = 194 \pm 33$ ns, $t_{\rm unfold} = 64 \pm 14$ ns, $t_{\rm diff} = 6 \pm 3$ ns. 106

3.11 On the top the MFPTs of the all possible contiguous folded substrings is shown: an entry in the triangular map represents the estimated MFPT necessary to form a chain folded fragment having length going from 1 (single residue) to the full length chain. The length of a substring gives the hierarchy level on the y axes of the maps. The fragment MFPTs are computed on the whole ensemble of strings along the time series. The maps provide an overview of the possible folding mechanisms. The kinetic map of the polyTHR_5GS appears homogeneous suggesting that folding takes place cooperatively. Conversely the map for 1pgb_AGT looks modular with respect to the elements of secondary structure. This suggests that folding may be non-cooperative: 1st hairpin and helix plus 2nd hairpin diffuse and collide into each other. From the maps only the hierarchy trees for 1pgb_AGT could be extrapolated (bottom right in figure). At the lowest hierarchical level R walkers start a random walk, namely that 1-fragments are assembled to gain the next status level of 2-fragments. The algorithm makes the walkers follow the maximal rate route, for instance two 1-fragments can assemble into two different 2-fragments, thus the algorithm picks that which maximizes the 2-fragment formation rate. The procedure is repeated for all the hierarchy levels until a tree is completed by reaching the full folded string that lies on the top of the tree. The algorithm finds the maximal rate tree associated to the map. On bottom left an interpretation of the possible folding pathways of 1pgb_AGT. 108

3.12 See caption of figure **??**. 110

3.13 The networks corresponding to the transition matrices extrapolated from the time series of 200 causal grouped SRA[4] mesostates for the proteins polyTHR_2GS and polyTHR_5GS (figure **??** (A) and (B) respectively) and RMSD[5.0] mesostates for 1pgb_AGT (current figure). The figures were realized with the program Tulip [Auber, 2003] and the visualization algorithm applied is the so called spring-embedder. The links are colored according to the values of the transition matrix: darker colors correspond to high transition probabilities while clearer colors to lower values. Accordingly the color of the nodes resemble the mean value of the in-going and out-going edges. Node sizes are chosen without a numerical criteria but only to facilitate the graph reading. Cluster of nodes are grouped together into basins according to their inter-connectivities. To some nodes or basins the corresponding ensemble of structures are represented with their global populations. The graph of polyTHR_2GS is essentially divided in two main phases, a helix phase whose weight is ~ 40 % and a triple-stranded β-sheet phase whose weight is ~ 45 % which is the folded state. Curl-like basins are also present in both N- and C-term configuration whose weight is about 1.5 %, an aspect that is due to the symmetry of the sequence. The graph for polyTHR_5GS is much more complex due to the proliferation of non-folded β structures that also play the role of kinetic traps. The helix phase is still present though its population is reduced to about 10 %. The folded basin is large and populated about 30 %. Many "exotic" β rich basins are present. Finally the graph for 1pgb_AGT depicted in the current figure is very heterogeneous although the folded basin is clearly detectable. Two unfolded basins, H1and H2 are well defined and populated and characterized by a long helix packed respectively with a double and triple stranded β-sheet. Basin percentages indicated on networks are indicative values. 111

3.14 A uni-dimesional free energy profile for the protein 1pgb_AGT from the causal grouped mesostates RMSD [5.0]. The reaction coordinate is the calculated equilibrium MFPT from any mesostate to the folded through the evolution of the Markov chain on the causal mesostates. They are the values $M_{j\rightarrow folded}$ of the extrapolated matrix of the MFPTs. The values in the y axes are stability ΔG values extrapolated from the main diagonal of the MFPT matrix. The values are reported for less than an additive constant. Two main minima separated by a barrier are evident. The main unfolded basin is separated from the folded though about 1 kcal/mol barrier. A far basin relaxing very slowly to the folded state represents fibrous states. 113

3.15 The reordered MFPT matrix $M^*_{i\rightarrow j}$ for the protein 1pgb_AGT based on causal grouped mesostates from RMSD[5.0]. An entry on the matrix gives the MFPT for the equilibrium transition $i \rightarrow j$. Horizontal bands are equilibrium transitions from all the is to a specific j. The index (i,j) are ordered from 1 fastest relaxation to the folded state to 200 slowest relaxation to the folded state. The folded basin is composed by the dense bands going from 1 to about 50. The fact that the bands are dense means the folded state can be reached from many gateways in about 150/200 ns, in particular the Markovian MFPTs analysis confirms that an overall folding free energy barrier separates the unfolded state from folded. Most kinetically far states are the fiber-like states that are mainly non-equilibrium states (the system visits them barely once and never returns there). The proper unfolded state is thus populated by helices in several combinations (helix bundles, α/β, etc.) rapidly exchanging between them. 114

7.1 The scientific production on the "protein folding" topic of the last 40 years. The black curve represents the total number of published articles per year: after a constant increase from 1973 to the end of the 80s a bursting increase characterize the beginning of the 90s, which is followed by a plateau. The blue curve is the mean number of citations of the papers published in a certain year which is calculated from the total number of citations per year divided by the total number of published papers per year. The red curve gives the number of citations corresponding to the most cited paper per year: peaks corresponds to outstanding papers that represent a breakthroughs in the field (some representative citations are included). The blue and red curves in the late tails are clearly affected by an incomplete statistics. Do the trends suggest an imminent decline of protein folding as an autonomous research topic or the coming of new breakthroughs? 167

List of Tables

2.1 $^{(*)}$ [kcal/mol]. Thermodynamic parameters: $\Delta E_i = \overline{E_i} - \overline{E}, T\Delta H_i = T(H_i - H), T\Delta S_i^b = T(S_i^{\mathrm{b}} - S^{\mathrm{b}})$ (see equation **??** for the definition of S^{b}), $\Delta G_i = \Delta E_i - T\Delta S_i^{\mathrm{b}}$. The list of the first 50 most populated mesostates of the GSGS for the coarse graining based on the strings of native contacts SNC[2] with all the thermodynamic quantities estimated according to the equations in the text. 29

2.2 $^{(*)}$ [kcal/mol]. Same as table **??** for the coarse graining based on the strings of secondary structure SSS[8]. 30

2.3 $^{(*)}$ [kcal/mol]. Same as table **??** for the coarse graining based on the strings of rotational states SRA[4]. The quantity Δh_i is the configurational entropy loss of a string considering all the contributes due to the string sites, as explained in section **??**. 31

2.4 $^{(*)}$ [kcal/mol]. Same as table **??** for the clustering based on Cα-RMSD with cutoff 1.5 Å. . 32

2.5 $^{(*)}$ [kcal/mol/K]. Overview of the entropies for all the descriptive methods adopted to coarse grain the conformation space of the GSGS. ΔH_{gain} is the information gain due to the coarse grain and $\mathcal{D}$ is its statistical disorder and $\mathcal{C}$ the complexity. 35

2.6 The parameters estimated from the log-normal fit (H^*/k_B, N^*, $\beta\sigma_G$) of the densities of mesostates and the exponent D estimated from a power law fit (eq. **??**) mainly for the string based mesostates. The number of stable mesostates r_c and their cumulative statistical weight P_{stable} are also reported in table. The r_c number should be regarded as an order of magnitude rather as an exact quantity. 40

2.7 The thermodynamic parameters defining the selected mesostates of figure **??**. $\Delta E_i = \overline{E_i} - \overline{E}$, $T\Delta S_i^b = T(S_i^{\mathrm{b}} - S^{\mathrm{b}})$ (see equation **??** for the definition of S^{b}), $\Delta G_i = \Delta E_i - T\Delta S_i^{\mathrm{b}}$. Positive values of ΔG_i correspond to unstable mesostates (with positive ΔE_i and low $T\Delta S_i^b$). Among the unstable mesostates the bad sampled ones can be found. Positive values of ΔG_i are distributed according to a Gaussian (see figure **??**). Mesostates with negative ΔG_i are stable and well sampled: they posses either low enthalpy or large vibrational entropy. 43

2.8 Triple and double exponential fitting parameters of the FPT distributions for different descriptions of the configurational space. 54

2.9 The MFPTs of selected mesostates to the folded mesostate taken from table **??** and figure **??**. 57

3.1 Secondary structure and mesoscopic dihedral propensities estimated using the DASSD database [Dayalan et al., 2006]. 88

3.2 The simplified protein sequences here studied: polyTHRs and 1pgb_AGT. GSGS and 1pgb sequences are reported with the secondary structure and SRA[4] strings that correspond to the simulation structure and the X-ray (see [Gallagher et al., 1994] for the latter) respectively. 88

3.3 The total number of sampled microstates M in the simulations; the total number N of mesoscopic strings SSS[8], SRA[4] and the total number of clusters found. 90

3.4 $^{(*)}$ [kcal/mol]. The list of the first 10 most populated SRA[4] mesostates from the polyTHR_xGS simulations with the thermodynamic quantities estimated according to the equations developed in chapter **??**. In particular ΔE_i is the effective mean energy difference per mesostate, $\sigma(E_i)$ is the standard deviation of $\overline{E_i}$, P_i is the population per mesostate, $T\Delta h_i$ is the configurational entropy loss per mesostate, $T\Delta S_i^{\mathrm{b}}$ is the internal entropy difference per mesostate and ΔG_i is the free energy difference per mesostate. . . 92

3.5 $^{(*)}$ [kcal/mol]. The list of the first 50 most populated SRA[4] mesostates from the 1pgb_AGT simulations at 330 K with their thermodynamic quantities. Strings in bold are those corresponding to a structural motif compatible with that of protein G. There are many strings sharing such a structural feature, notably these strings have a negative configurational entropy loss $T\Delta h_i$ which means they are promoted by enthalpy. 97

3.6 $^{(*)}$ [kcal/mol]. The list of the first 20 most populated RMSD[5.0] clusters from the 1pgb_AGT simulations at 330 K with their thermodynamic quantities and structural representations. Many clusters correspond to the structural topology of protein G, in particular the clusters 1 (which is also shown in figure **??** (A) with its cluster center), 4, 10, 11, 17 and many others. There are also many clusters having the native secondary structure but an incorrect topology, for instance either one or two hairpins misplaced. These clusters are for example the 2, 3, 9, 13, 14, 18 among those shown in figure. Other interesting clusters are those having the β-sheet formed but the helix either in a coil or β structure: see for instance cluster 12 and 20. 98

Acknowledgments

First of all I would like to thank Amedeo Caflisch for his support, his open mindset, his scientific supervision and for the possibility he has given to me to accomplish my PhD. Not always we have been on the same wavelength but we both are difficult characters and at the end the reasonability and the reciprocal esteem has always prevailed. I'd like to thank Emanulele Paci for giving me the possibility to come to Zürich to begin this experience. I especially thank him for his scientific supervision and moral support of my first 3 years of PhD here in Zürich and for the 9 month at the Institute of molecular biophysics of the University of Leeds UK, where we both moved at the end of 2004 and where I was visiting student. Thanks to Ben Schuler of whom I like the intellectual originality and his scientific feedbacks. I would like to thank Riccardo Pellarin for his friendship and his ability to help me. We enjoy discussing about science since we were both naive students in theoretical physics in Rome. He has been irreplaceable and I won't thank him enough for his help for my work. I thank Andrea Cavalli for his friendship, his scientific intelligence, his eclectic pragmatism, his help. He was the first guy who talked to me when I joined the Caflish's group. Since then we didn't stop talking about everything. Then I want to thank Marco Cecchini and Francesco Rao for their open friendship, Francesco is now father and Marco soon will. I thank Gian Gaetano Tartaglia for his friendship during his stay here in Zürich. I thank Beatrice Paoli, Andrea Prunotto, Pietro Alfarano and Marino Convertino for their affection and warm freshness. I thank Stefanie Muff for the many discussions we have had about work. I thank Gianluca Interlandi for his being a very spontaneous person in a world of good looking soldiers. I thank Fabio Parmegiani for his friendship and the interest he has shown for my work. I thank Dariusz Ekonomiuk, Ran Friedman,Danzhi Huang, François Marchand, Philipp Schütz and Ting Zhou for the warm and friendly atmosphere they always gave to me. Then I thank Alex Abebe, Shaheen Ahmed, Rainer Böckmann, Raffaele Curcio, Fabian Dey, Joerg Gsponer, Urs Haberthuer, Peter Kolb, Nicolas Majeux, Michele Seeber, Gianni Settanni for the many discussions we have had and for their sincere companionship. Many thanks to Jean Claude Tomasina and Christiane Gujan for their assistance and sincere kindness. Then I'd like to thank the people of Leeds with whom I shared important moments. I'd like to thank Manlio Tassieri, his wife Loredana and his little daughter for their sincere and warm friendship. I'd like to thank Horacio Montes De Oca, Bernardo Perez, Natalia and Malena for letting me dreaming about Mexico and Latin America, while I was in the dark winter of north England. I'd like to thank Dan for his kindness and my flat-mates John, Jan, Jean Claude and Malidi. I would like to thank Davide & Beatrice, Anna & Axel and their son, Alexander, Marcello, Nathalie, Andrea, Marco, Pierluigi,... etc., etc. I'd like to say thanks to Svava, my love, with whom I discovered the naturalness of sharing joys and pains. Thank you, I'm not sure I would have made it without you. Thanks to my parents, my brother, his wife Eliana, my two crazy nieces Arianna and Gloria and to all my beloved friends of ever in Rome just for existing and supporting me always. Finally thanks to the reader, who I hope will forgive me if here I have maybe too often indulged in some intellectual nonchalance. ♠

Enrico **GUARNERA**

Date and place of birth: 4th February 1973, Rome (Italy)

Address: Glattstegweg 38
8051 Zürich, Switzerland

Email: guarnera@bioc.uzh.ch

Nationality: Italian

EDUCATION AND STUDIES

since 2002: **PhD studies** in the group of Prof. Dr. Amedeo Caflisch, Department of Biochemistry, University of Zürich, Switzerland, under the supervisions Prof. Dr. Amedeo Caflisch and Dr. Emanuele Paci (University of Leeds, UK)
Thesis title: "*Theoretical and computational strategies for the study of protein folding mechanisms*"

2000: **Master Degree in Theoretical Physics** passed the 29th November 2000 at University of Rome III, Italy.
Final score: 110/110.
Thesis title: "*Super symmetric models of fermion masses in U(1) theories*." Supervisor Prof. Dr. Guido Altarelli (CERN Geneva, Roma III)

1994 - 2000: **Undergraduate studies in Theoretical Physics** at University of Roma III.

1993: Science-technology **High School Degree** at "ITIS F. Severi" of Rome passed with score 60/60.

TRAINING PERIODS

2004 - 2005: Visiting Student at the Institute of Molecular Biophysics and Physics Astronomy department of the University of Leeds, UK, under the supervision of Dr. Emanuele Paci.

LANGUAGES

Italian: mother tongue
English: fluent, written and spoken
Spanish: fluent, written and spoken
German: basic knowledge